Degrees of Freedom Tests for Smoothing Splines

Eva Cantoni and **Trevor Hastie**

Statistics Department
Stanford University
Sequoia Hall
390 Serra Mall
Stanford, CA 94305

June 2000

Abstract

When using smoothing splines to estimate a function, the user faces the problem of choosing the smoothing parameter. Several techniques are available to select this parameter according to certain optimality criteria. This involves solving an optimization problem. Here, we choose a different point view and we propose a technique to choose between two alternatives (e.g. allowing for two different levels of degrees of freedom). The problem is addressed in the framework of a mixed-effects model, a likelihood-ratio-type test statistic is proposed, and its distribution is derived. A test of linearity follows directly. We then extend this idea to additive models where it provides a more attractive alternative than multiparameter optimization, and where it gives exact distributional results that can be used in a analysis-of-deviance type of approach. Examples on real data, and a simulation study of level and power complete the article.

1 Motivation

Selecting a value for the smoothing parameter, or equivalently the effective degrees of freedom, is a common problem in nonparametric regression. Consider the left panel of Figure 3; one could ask the question: does the solid line describe the relationship between `Row` and `Total yield` well enough or is a more flexible fit (like the dashed line) needed? The aim of this paper is to provide a test statistic to choose between two such alternatives.

We consider the model

$$y_i = f(x_i) + \epsilon_i, \tag{1}$$

for each individual i of a sample of size n. We focus our attention upon the estimation of the function f by smoothing splines. It is well known that the smoothing spline fit is computed as a linear transformation of the vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ and that the fitted values are $\hat{\mathbf{y}}_\lambda = (I + \lambda K)^{-1}\mathbf{y} = S_\lambda \mathbf{y}$ (see Green and Silverman, 1994, and Appendix A.1 for more details). The parameter λ controls the smoothness of the fit, which can also be defined through the effective degrees of freedom. As in Hastie and Tibshirani (1990), we define the effective degrees of freedom of a smoothing spline fit by

$$\mathsf{df} = \mathsf{Tr}(S_\lambda) = \sum_{i=1}^{n} \frac{1}{1 + \lambda d_i}, \tag{2}$$

where d_i are the eigenvalues of the matrix K (the quadratic form $\hat{\mathbf{y}}_\lambda^T K \hat{\mathbf{y}}_\lambda$ computes the roughness of the fitted function). There is a strictly monotone relationship between λ and the number of degrees of freedom, allowing us to work with this latter notion, which ties in gracefully with parametric linear modeling concepts.

Classical approaches of degrees of freedom selection have considered the optimization of some optimality criteria (e.g. estimators of the mean squared error). This is the case for cross-validation or Mallows's C_p, for example. Another common approach for determining the value of the smoothing parameter consists of deriving analytical expressions for the mean squared error, from which the optimal value of the parameter is obtained. This optimal value usually depends on the underlying unknown function and therefore this approach is called a "plug-in" approach, because of its need of a pilot estimator.

Here we take a different point view: we would like to construct a test statistic to choose between two predefined alternatives, much like a parametric modeling. Tests like this are quite important, in particular in view of the extension to additive models. In fact, it is usual practice to use such tests to build additive models (see, for example, Hastie and Tibshirani, 1990). This approach avoids the optimization of a multidimensional criterion. An additive model in 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: *out, linear,* $4df$ *or* $8df$, and the techniques discussed in this paper allow us to test hypotheses to choose between them.

To pursue our goal, we consider the mixed-effects framework for smoothing splines. This approach has the advantage of avoiding the problem of biased estimators, well known in nonparametric regression.

Section 2 gives the basics of the mixed-effects approach to smoothing splines. The test statistic we are interested in is derived in Section 3. A real example is worked out in Section 4, which is followed by Section 5 devoted to a simulation study of the properties of the likelihood-ratio statistic. Finally, Section 6 considers the extension to additive models. The article is completed with a concluding section and with technical details collected in the appendix.

2 Mixed-effects setting for smoothing splines

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

$$\mathbf{Y} \quad = \quad X\boldsymbol{\beta} + Z\mathbf{u} + \boldsymbol{\epsilon}, \tag{3}$$

made up of a *linear* fixed effect $X\boldsymbol{\beta}$, a *nonlinear* random effect $Z\mathbf{u}$ and an independent and unstructured error term. In detail,

- $X\boldsymbol{\beta}$: $X = [\mathbf{1}\ \mathbf{x}] \in \mathbb{R}^{n\times 2}$, where $\mathbf{x} = (x_1, \ldots, x_n)^T$ is the vector of predictor values, and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ are unknown parameters.
- $Z\mathbf{u}$: $Z = Z(\mathbf{x}) \in \mathbb{R}^{n\times(n-2)}$ is a matrix representing nonlinear functions of x (and the columns of Z are orthogonal to the columns of X).

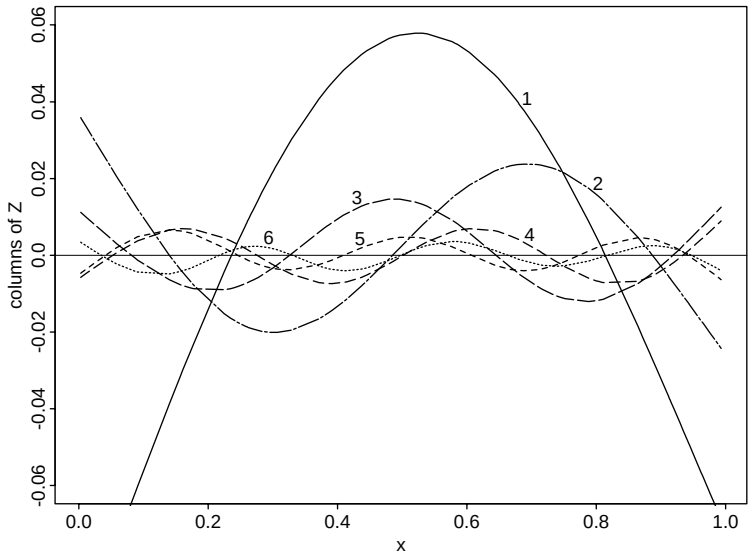

Figure 1: *Six first columns of $Z(x)$ based on a sample* $\mathbf{x}$ *of size* $n = 100$ *drawn from a uniform distribution.*

The successive columns of Z are increasingly rough (as measured by the quadratic penalty matrix K), and are scaled to have decreasing Euclidean norm. Figure 1 shows the six first columns of $Z(x)$ based on a sample $\mathbf{x}$ of size $n = 100$ drawn from a uniform distribution. $\mathbf{u} \sim \mathcal{N}(0, \tau^2 I_{n-2})$ 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.

- $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$ is the error term, independent of $\mathbf{u}$.

Further technical details on the matrix Z are given in Appendix A.1.

The BLUP predictors for $\boldsymbol{\beta}$ and $\mathbf{u}$ of model (3) satisfy the following equations (see Robinson, 1991 and the original reference by Henderson, 1950):

$$\begin{aligned} (X^T X)\hat{\boldsymbol{\beta}} &= X^T(\mathbf{y} - Z\hat{\mathbf{u}}) \\ (Z^T Z + \lambda I_{n-2})\hat{\mathbf{u}} &= Z^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}), \end{aligned}$$

where $\lambda = \sigma^2/\tau^2$ is assumed known. These equations are obtained by maximization of the joint density of $\mathbf{y}$ and $\mathbf{u}$ with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$ under the normality assumptions, see Henderson (1973) and Robinson (1991).

Taking into account the orthogonality $X^T Z = Z^T X = 0$ (see Appendix A.1), it follows that the predictors for $\boldsymbol{\beta}$ and $\mathbf{u}$ are

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \quad (4)$$

$$\hat{\mathbf{u}} = (\lambda I_{n-2} + Z^T Z)^{-1} Z^T \mathbf{y}. \quad (5)$$

BLUP stands for *best linear unbiased predictors*, that is the best predictors in term of mean squared error among the class of linear unbiased functions of the data. Unbiased refers here to the property that the average value of the estimate is equal to the average value of the quantity being estimated, that is $E(\hat{\mathbf{u}}) = E(\mathbf{u})$. Note that it is common practice to make the distinction between estimators (for fixed effects) and predictors (for random effects).

We show in Appendix A.1 that the fitted values obtained with (4) and (5), namely $\hat{\mathbf{y}}_\lambda = X\hat{\boldsymbol{\beta}} + Z\hat{\mathbf{u}}$, are the same as $\hat{\mathbf{y}}_\lambda = S_\lambda \mathbf{y}$. This shows (Speed 1991) that the BLUP estimates obtained for this particular mixed-effects model are identical to the smoothing spline obtained by penalized least squares.

So far we have assumed τ^2 and σ^2 to be known; the (marginal) Gaussian likelihood for $\mathbf{Y}$ also allows inference for these parameters, or the noise-to-signal ratio λ. In our problem the marginal distribution of $\mathbf{Y}$ is a $\mathcal{N}(X\boldsymbol{\beta}, \sigma^2(I_n + 1/\lambda ZZ^T))$ distribution. For example, the value of the smoothing parameter λ could be estimated by maximum likelihood, see Wecker and Ansley (1983) and Wahba (1990). This approach is called the generalized maximum likelihood criterion and is equivalent to the maximum likelihood estimation derived from model (3). In this paper, we pursue a different idea which consists of deriving a likelihood-ratio type test to choose between two alternatives.

There is a fully Bayesian approach which is equivalent to this mixed-effects model. In this case $\boldsymbol{\beta}$ is also random, having a uniform improper prior. The posterior means of $\boldsymbol{\beta}$ and $\mathbf{u}$ are identical to (4) and (5). In this context inference on the variance parameters from the marginal likelihood is referred to as *Empirical Bayes*, and is identical to the inference in the mixed-effects framework. Differences would occur if a proper prior were assumed for $\boldsymbol{\beta}$.

3 A likelihood ratio test statistic

Suppose that σ is known. The inference on the smoothing parameter can be carried out on the parameter τ of the marginal distribution. So we consider the test of the null hypothesis $H_0 : \tau = \tau_0$ versus the alternative hypothesis

$H_A : \tau = \tau_1 > \tau_0$. This corresponds to the test of $H_0 : \lambda = \lambda_0$ versus $H_A : \lambda = \lambda_1 < \lambda_0$, or equivalently $H_0 : \mathsf{df} = \mathsf{df}_0$ against $H_A : \mathsf{df} = \mathsf{df}_1 > \mathsf{df}_0$.

Denote by $l_\tau(\mathbf{y})$ the density associated to the marginal distribution of $\mathbf{Y}$, namely the density of a $\mathcal{N}(X\boldsymbol{\beta}, \sigma^2(I_n + 1/\lambda ZZ^T))$ distribution. We then consider the log likelihood-ratio statistic based on this distribution:

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

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

We can define a test statistic by

$$\begin{aligned}
T &= \mathbf{y}^T \big((I_n + 1/\lambda_0 ZZ^T)^{-1} - (I_n + 1/\lambda_1 ZZ^T)^{-1}\big)\mathbf{y} \\
&= \mathbf{y}^T \big(S_{\lambda_1} - S_{\lambda_0}\big)\mathbf{y} \\
&= \mathbf{y}^T \big(\hat{\mathbf{y}}_{\lambda_1} - \hat{\mathbf{y}}_{\lambda_0}\big), \qquad (6)
\end{aligned}$$

where $\hat{\mathbf{y}}_\lambda = S_\lambda \mathbf{y}$ are the fitted values and where the middle expression holds because of relationship (21) in Appendix A.1.

The distribution of T depends on σ^2, see Section 3.1 below. One can either plug-in a reliable estimate, or – better – consider an F-type statistic like

$$F = \frac{\mathbf{y}^T \big(S_{\lambda_1} - S_{\lambda_0}\big)\mathbf{y}}{\mathbf{y}^T (I - S_{\tilde{\lambda}})\mathbf{y}} = \frac{\mathbf{y}^T \big(\hat{\mathbf{y}}_{\lambda_1} - \hat{\mathbf{y}}_{\lambda_0}\big)}{\mathbf{y}^T \big(\mathbf{y} - \hat{\mathbf{y}}_{\tilde{\lambda}}\big)}, \qquad (7)$$

for some value $\tilde{\lambda}$ of the smoothing parameter. We will discuss the issues related to the choice of $\tilde{\lambda}$ in Section 3.1 along with the distribution of the statistic F.

The generalization to the test of composite hypotheses of the form $H_0 : \lambda = \lambda_0$ against $H_A : \lambda > \lambda_0$ (unspecified) can be obtained by estimating λ (and σ^2) using maximum likelihood under model (3) and use it instead of λ_1.

For projection operators (like linear regression), statistic (7) is equivalent to the statistic that would compare the sums of squared residuals of the fits, because in this case $\mathbf{y}^T(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}})$. In fact, the heuristic approach used by Hastie and Tibshirani (1990) and Chambers and Hastie (1991) 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

$$\Delta = \frac{\big((\mathbf{y} - \hat{\mathbf{y}}_{\lambda_0})^T(\mathbf{y} - \hat{\mathbf{y}}_{\lambda_0}) - (\mathbf{y} - \hat{\mathbf{y}}_{\lambda_1})^T(\mathbf{y} - \hat{\mathbf{y}}_{\lambda_1})\big)/(\nu_1 - \nu_0)}{(\mathbf{y} - \hat{\mathbf{y}}_{\lambda_1})^T(\mathbf{y} - \hat{\mathbf{y}}_{\lambda_1})/(n - \nu_1)}, \qquad (8)$$

which is approximated by an $F_{\nu_1-\nu_0,n-\nu_1}$ distribution with $\nu_i = \mathsf{df}_i = \mathsf{Tr}(2S_{\lambda_i} - S_{\lambda_i}S_{\lambda_i}^T)$ for $i = 0, 1$. This means that it is assumed that the numerator and the denominator in (8) are approximated by a $\chi^2_{\nu_1-\nu_0}$ and a $\chi^2_{n-\nu_1}$ variable respectively. A further and computationally less expensive approximation consists of taking $\nu_i \simeq \mathsf{Tr}(S_{\lambda_i})$. For instance, the implementation of this heuristic procedure in the statistical package S-PLUS makes use of this latter approximation.

There are several levels of approximations involved in this procedure, and we would like to investigate them. The approach relies on a model of the form

$$y_i = f(x_i) + \epsilon_i, \qquad (9)$$

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 (8) is a linear combination of noncentered χ^2_1 variables. Hence, saying that the numerator is $\chi^2_{\nu_1-\nu_0}$ distributed accounts for two different sources of approximation. First of all, function estimates by smoothing techniques based on model (9) are almost always biased and this bias, which depends on the true underlying function f, is neglected when using statistics (8). Secondly, the weighted χ^2_1 combination is approximated by a unique χ^2 variable. The same comments apply to the denominator in (8). In addition, the F-approximation to the ratio-statistic Δ does not take into account the dependence between its numerator and its denominator.

All these reasons make our new exact procedure very attractive. In particular, although the procedure using Δ could be improved by refining the distributional properties (see, for example, Hastie and Tibshirani, 1990, p. 66-67), the bias problem will still be present. This problem is avoided with our Bayesian procedure.

3.1 Distribution of the test statistic under H_0

We first look at the numerator of statistic (7), which is in fact statistic T as defined in (6). The distributional properties of this statistic are easily

established, because it turns out to be a quadratic form in normal variables. We have that under H_0

$$T = \sigma^2 \sum_{i=3}^{n} c_i z_i^2, \tag{10}$$

with $c_i = 1 - \frac{d_i + 1/\lambda_0}{d_i + 1/\lambda_1}$, and z_i being standard normal variables, see Appendix A.2 for details. It has to be noticed that, consistently with the results on quadratic forms in normal variables, the c_i's are in fact the eigenvalues of $(I_n + 1/\lambda_0 ZZ^T)(S_{\lambda_1} - S_{\lambda_0})$.

Also under H_0, the variable on the denominator is distributed according to a $\sigma^2 \sum_{i=3}^{n} b_i \chi^2_{1(i)}$ distribution, where the b_i are the eigenvalues of $(I_n + 1/\lambda_0 ZZ^T)(I_n - S_{\tilde{\lambda}})$ and are equal to $\frac{d_i + 1/\lambda_0}{d_i + 1/\tilde{\lambda}}$. The proof of this result goes through the same lines as the proof of the distribution of T and is omitted.

Both these test statistics are distributed according to a linear combination of χ^2 variables with one degree of freedom. These kinds of distributions have been well studied (see, for example Chapter 29 in Johnson and Kotz, 1970), and algorithms are available to compute probabilities accordingly, e.g. Davies (1980) and Farebrother (1990).

Although, the distribution of its numerator and its denominator are known, the distribution of F itself does not belong to a known family. One could consider approximating each of the two linear combinations of χ^2_1 variables by unique χ^2 variable that matches the first moment, that is 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 F is approximately $(\sum c_i / \sum b_i) F_{\sum c_i, \sum b_i}$, with $\sum c_i = \mathsf{Tr}((I_n + 1/\lambda_0 ZZ^T)(S_{\lambda_1} - S_{\lambda_0}))$ and $\sum b_i = \mathsf{Tr}((I_n + 1/\lambda_0 ZZ^T)(I_n - S_{\tilde{\lambda}}))$. Both these traces have more computational appealing expressions, see Appendix A.3.

To guarantee approximate independence between the numerator and the denominator of the F statistic in (7) one has to choose a "small" value $\tilde{\lambda}$ of the smoothing parameter. One example of the 50 larger weights b_i and c_i for a uniformly distributed design of size $n = 100$ on the interval $(0, 1)$ and for a value of $\tilde{\lambda}$ corresponding to 20 degrees of freedom is represented in Figure 2. As one can see from this plot, by trying to achieve near independence, the weights b_i of the χ^2 combination of the denominator in (7) are spread out, which could potentially have harmful consequences on the accuracy of the approximation by a unique χ^2 variable that matches the first moment as

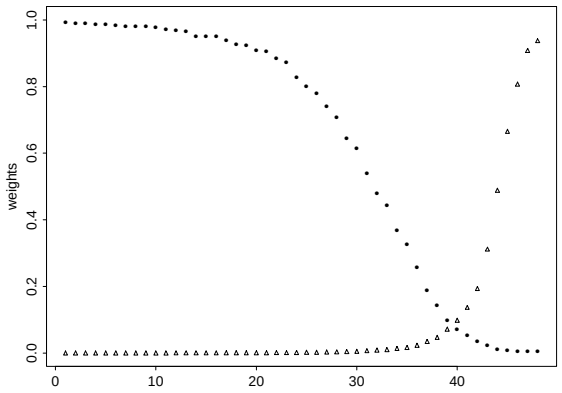

Figure 2: *50 larger weights b_i (circles) and c_i (triangles) for a uniformly distributed design. Sample size is $n = 100$.*

suggested above. A better approximation can be obtained through a two-moment correction.

However, one is usually interested in obtaining p-values, which can be computed exactly much more easily by slightly modifying the problem. In fact, the p-value $P(F > f_{obs})$ – for the value f_{obs} that the statistic (7) takes on the dataset – can be rewritten as

$$P\Big(\mathbf{y}^T(S_{\lambda_1} - S_{\lambda_0} - f_{obs}(I - S_{\tilde{\lambda}}))\mathbf{y} > 0\Big), \tag{11}$$

which involves again a quadratic form in normal variables. The distribution of $R = \mathbf{y}^T(S_{\lambda_1} - S_{\lambda_0} - f_{obs}(I - S_{\tilde{\lambda}}))\mathbf{y}$ is also easily established, because

$$\begin{aligned} R &= \mathbf{y}^T\Big(S_{\lambda_1} - S_{\lambda_0} - f_{obs}(I - S_{\tilde{\lambda}})\Big)\mathbf{y} && (12)\\ &= \sum_{i=3}^{n} e_i z_i^2, && (13) \end{aligned}$$

with $e_i = c_i - f_{obs}b_i = 1 - \frac{d_i + 1/\lambda_0}{d_i + 1/\lambda_1} - f_{obs}b_i$, or – otherwise stated – the e_i are the eigenvalues of $(I_n + 1/\lambda_0 ZZ^T)(S_{\lambda_1} - S_{\lambda_0} - f_{obs}(I - S_{\tilde{\lambda}}))$.

The distribution of (12) is again a linear combination of χ_1^2 variables and can be computed with the algorithms previously discussed. Result (13) is obtained in the same way as the distribution of the statistic (6), and we do not give the details.

The approach by (11) is appealing because it has the two main properties we were looking for. It is scale invariant and is not affected by the lack of independence between the variable of the numerator and the denominator in (7). This approach has been used by other authors, see for instance Azzalini and Bowman (1993).

Since with this approach we do not need to care about the independence between the numerator and the denominator in (7), the choice of $\tilde{\lambda}$ is of no concern. We suggest to use $\tilde{\lambda} = \lambda_1$ to avoid additional computations. In this case, we have that $e_i = 1 - (1 - f_{obs})\frac{d_i + 1/\lambda_0}{d_i + 1/\lambda_1}$.

A test of linearity follows directly from the test statistic F. In fact, testing linearity corresponds to the test of $H_0 : \lambda = \lambda_0 = \infty$ versus $H_A : \lambda = \lambda_1 < \infty$. We obtain in this way a competitor to the linearity test of Azzalini and Bowman (1993).

3.2 Computational aspects

The whole exact procedure we propose is rather computational expensive, that is, of complexity $O(n^3)$, although statistics (7) and (12) themselves can be computed in linear time (because fitted values can be obtained in $O(n)$ steps). However, the speed of nowadays computers is such that samples up to moderate sample sizes can be handled in very few seconds. Thanks to expressions (22) and (23), which require only the trace of the smoother matrix, the approximation to the distribution of (7) by an F distribution involves only $O(n)$ computations. If this approximation will prove accurate, it substantially reduces the computational burden of the procedure.

The procedure of Hastie and Tibshirani (1990) and Chambers and Hastie (1991) based on sums of squared residuals is of order $O(n^2)$ when the degrees of freedom are computed as $\nu_i = \mathsf{Tr}(2S_{\lambda_i} - S_{\lambda_i} S_{\lambda_i}^T)$, even if the statistic Δ itself can be computed in linear time. The computational price to pay goes down to $O(n)$ when the approximated degrees of freedom $\nu_i \simeq \mathsf{Tr}(S_{\lambda_i})$ are used instead.

4 Example

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

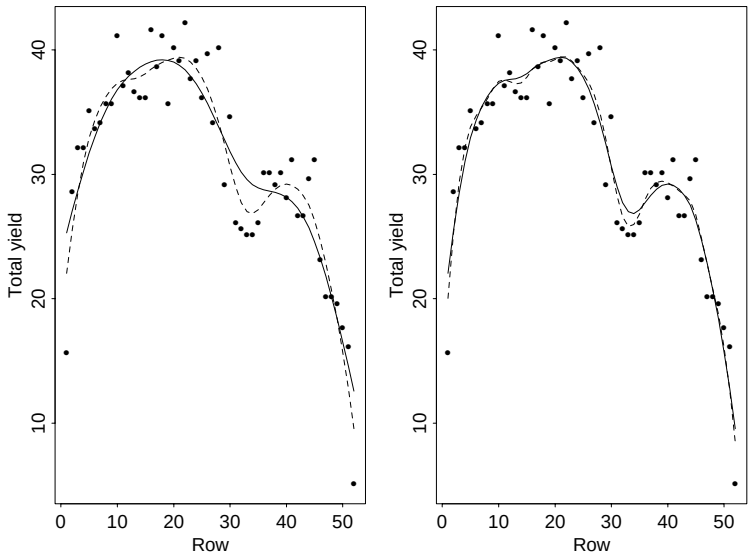

Figure 3: *Fit comparison on the vineyard data set. Left panel compares a spline fit with 6 df (solid line) to a spline fit with 10 df (dashed line). The right panel compares a fit with 10 df (solid line) to a fit with 14 df (dashed line).*

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 data set correspond to the sum of the yields of the harvests in 1989, 1990 and 1991. The yield is measured in number of *lugs*, a lug being one basket used to carry the harvest grapes.

Suppose we want to choose the degree of freedom in the set $\{6\mathsf{df}, 10\mathsf{df}, 14\mathsf{df}\}$. We perform two tests. The first one corresponds to the left panel of Figure 3, 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 (7), the null hypothesis is clearly rejected (p-value < 0.002). The right panel of Figure 3 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 of 20%), 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

As we have seen in Section 3.1, the distribution of the test statistic (7) is not known exactly, although we can compute exact p-values. Moreover, an approximation has been worked out. In the present section, we conduct a small simulation study to have a perception of the sensitivity of the approximation with respect to the exact result. We also compare our technique to the heuristic procedure based on sum of squared residuals discussed above.

We consider the simulation setting defined by the null model obtained by

$$\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2(I_n + 1/\lambda_0 ZZ^T)), \tag{14}$$

with $\boldsymbol{\beta} = (1, 5)^T$, $\sigma^2 = 0.5^2$, and λ_0 corresponding to 4 degrees of freedom. The alternative hypothesis considers a smoothing parameter λ_1 corresponding to 7 degrees of freedom. Moreover, $X = [\mathbf{1}\ \mathbf{x}]$, where $\mathbf{x}$ is generated from a uniform distribution at the beginning of the simulation.

We will compare the following situations.

(I) $F \sim (\sum c_i / \sum b_i) F_{\sum c_i, \sum b_i}$ with $\sum c_i$ and $\sum b_i$ according to (22) and (23) and $\tilde{\lambda}$ corresponding to 20 df.

(II) $R \sim \sum e_i \chi^2_{1(i)}$.

(III) $\Delta \sim F_{\nu_1 - \nu_0, n - \nu_1}$, with $\nu_i = \mathsf{Tr}(2S_{\lambda_i} - S_{\lambda_i} S_{\lambda_i}^T)$.

(IV) $\Delta \sim F_{\nu_1 - \nu_0, n - \nu_1}$, with $\nu_i = \mathsf{Tr}(S_{\lambda_i})$.

Cases (I) and (II) uses our procedure based on statistic (7) with approximate F and exact distribution respectively. Case (III) is the procedure based on statistic (8) with the matching degrees of freedom, whereas case (IV) is the same procedure but with approximated degrees of freedom (as implemented in S-PLUS, for example).

5000 simulations were ran with sample sizes $n = 40$ and $n = 100$. The precision in Davies's algorithm for the computation of the exact result has been set to 0.0001.

5.1 P-values and level

We first show some QQ-plots of the p-value of each techniques against the uniform distribution (Figure 4). The top panels are for the sample size $n = 40$

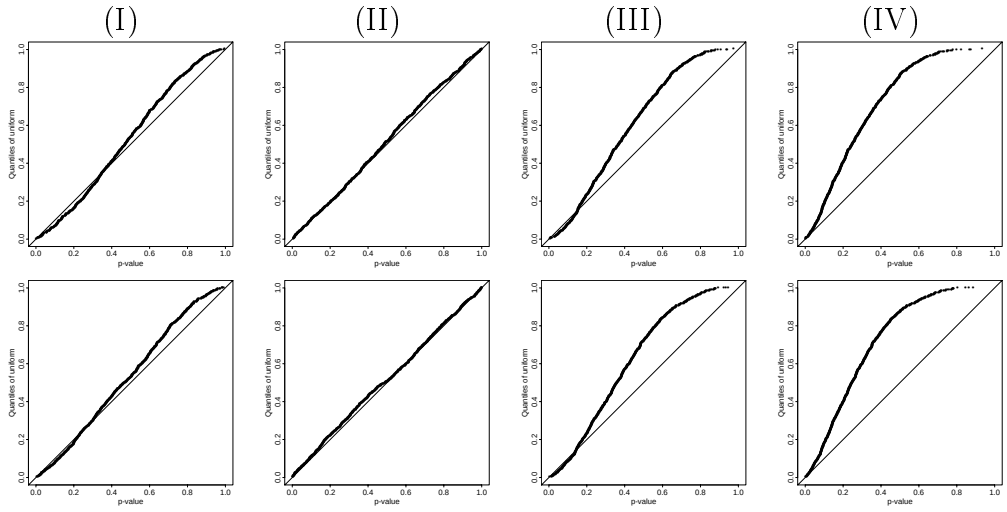

Figure 4: *QQ-plots of the simulated p-values of the test statistic* (7) *against the quantiles of the uniform distribution for cases (I), (II), (III) and (IV) when testing* $H_0 : df_0 = 4$ *versus* $H_A : df_1 = 7$. *Sample size is* $n = 40$ *for the top panels and* $n = 100$ *for the bottom panels.*

and the bottom panel are for $n = 100$. These QQ-plots show that approximation (I) to the distribution of (7) 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 right tail, which is fortunately of minor interest in a testing procedure. The use of the test statistic Δ, cases (III) and (IV), gives rise to even worse results than approximation (I). In both cases, the distribution of the p-value is far from the uniform target and is clearly skewed. We notice that this skewness is even more sensible in case (IV). Moreover, values of the p-value near to 1 never appear in the simulation for these two cases. As theoretically expected, the exact distribution yields uniformly distributed p-values, which also certifies that the numerical algorithm is accurate. For all the techniques, there does not seem to be any difference between small and large sample sizes.

Another way of assessing the quality of approaches (I), (II), (III) and (IV) is by looking at the actual level of the test with respect to some usual nominal levels, say $1, 5$ and 10%. For this purpose, we looked at the p-values previously simulated under H_0 and have counted how many times, out of the 1000 simulated cases, the null hypothesis was rejected. Table 1 displays these proportions (actual levels) as a function of the nominal level, the sample size

$\mathsf{df}_0 = 4, \mathsf{df}_1 = 7$	Nominal 1%		Nominal 5%		Nominal 10%	
Setting	$n = 40$	$n = 100$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
(I)	0.0038	0.0038	0.0296	0.0318	0.0692	0.0774
(II)	0.0086	0.0108	0.0452	0.0548	0.0952	0.1040
(III)	0.0028	0.0036	0.0284	0.0298	0.0720	0.0840
(IV)	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

Table 1: *Actual level of the test statistic* (7) *under techniques (I), (II), (III) and (IV) when testing* $H_0 : \mathsf{df}_0 = 4$ *versus* $H_A : \mathsf{df}_1 = 7$. *The last line of the table gives the standard deviations of the level estimation (do not depend on* n*).*

and the technique. The last line of the table gives the standard deviations of the level estimation, which do not depend on n.

As expected theoretically, the level computed by the exact distribution (II) behaves well: the actual level matches the nominal level for both sample sizes. The nominal level is always covered by the confidence interval constructed with $\pm$ twice the standard deviation.

For approximation (I), the actual level ranges between 30% and 75% of the nominal level, no matter the sample size. The behavior of the actual level is particularly bad on the extreme of the distribution (c.f. results at the 1% nominal level). In case (III), the results are similar – or probably slightly worse – to those of case (I), with an actual level in a range of $10-82\%$ of the nominal level. Nevertheless, both tests under (I) and (III) are conservative. This is not the case for the approach by (IV), which is clearly not on target, in particular at the 5% and 10% nominal level. In a certain sense, this is of no surprise in view of Figure 4 and due to the accumulation of different sources of approximation. The confidence interval constructed with $\pm$ twice the standard deviation do not cover the true nominal level in these cases, except for case (IV) and $n = 100$.

As a general conclusion we can say that the results of this simulation show that the F approximation to statistic (7) can be conservative, but it has the advantage of being inexpensive. The exact result is more computationally expensive, but the gain in accuracy is high. Therefore the use of the accurate exact distribution in the modified problem (13) is strongly recommended when computationally feasible. The heuristic procedure has a worse behavior,

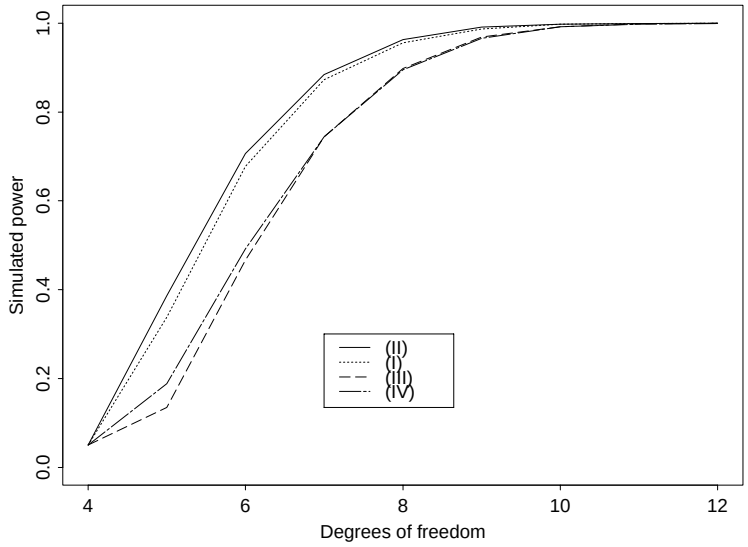

Figure 5: *Simulated power of the test statistics for cases (I), (II), (III) and (IV) (corrected to ensure a level of 5%).*

as bad as to be nonconservative in case (IV). Both approaches (I) and (II) outperform the results obtained by (III) and (IV).

5.2 Power

In the same setting as for the level simulation above, we also conducted a power study by simulation for a sequence of alternatives: $\mathsf{df}_1 = 5, \ldots, 12$ when $n = 40$. We simulated observations from model (14) with λ_0 replaced by λ_1.

The power curves (corrected to ensure a level of 5%) for the four situations (I), (II), (III) and (IV) are displayed in Figure 5. These results show an overall superiority of the test statistics (II) developed in this paper over the other approaches we considered. The power of the F-approximation of our test statistic – (I) – is almost as high as the power of the exact test. It has to be noticed that for a same (or less) computational cost, and considering that the error on the level was of the same magnitude, the F-approximation (I) does a better job in terms of power than the approaches (III) and (IV).

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 optimization of an optimality criterion over a p-dimensional space. From now on, we consider a p-dimensional additive model of the form

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \tag{15}$$

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

$$\mathbf{Y} = \beta_0 \mathbf{1} + X\boldsymbol{\beta} + \sum_{j=1}^{p} Z_j(\mathbf{x}_j)\mathbf{u}_j + \boldsymbol{\epsilon},$$

with $\mathbf{1} = (1, \ldots, 1)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $X = [\mathbf{x}_1 \ldots \mathbf{x}_p]$. As a result, the marginal distribution of $\mathbf{Y}$ is $\mathcal{N}(\beta_0\mathbf{1} + X\boldsymbol{\beta}, \sigma^2(I_n + \sum_j 1/\lambda_j Z_j Z_j^T))$.

Additive models are usually fitted via the *backfitting algorithm*, and – at convergence – the solution can be written as $\hat{\mathbf{y}}_{\boldsymbol{\lambda}} = R_{\boldsymbol{\lambda}}\mathbf{y} = \sum_j R_{\lambda_j}\mathbf{y}$, for a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$ of smoothing parameters. The degrees of freedom of the overall fit is $\mathsf{df} = \mathsf{Tr}(R_{\boldsymbol{\lambda}})$, which can be decomposed in its p components $\mathsf{Tr}(R_{\lambda_j})$. But this definition of the single component degree of freedom is not attractive from the computational point of view (see the discussion in Hastie and Tibshirani, 1990, p. 128-129). We will use instead the following definition

$$\mathsf{df}_j = \mathsf{Tr}(S_{\lambda_j}) - 1, \tag{16}$$

where S_{λ_j} is the smoother matrix obtained when fitting by smoothing spline the j-th predictor only. The subtraction of one is due to the global constant term isolated in model (15).

To perform a test on the degrees of freedom of the k-th component of model (15), we can define a test statistic by analogy with (7). Formally, we would like to test the hypothesis $H_0 : \mathsf{df}_k = \mathsf{df}_{k,0}$ (corresponding to $\lambda_k = \lambda_{k,0}$) versus $H_A : \mathsf{df}_k = \mathsf{df}_{k,1} > \mathsf{df}_{k,0}$ (corresponding to $\lambda_k = \lambda_{k,1}$). It is assumed that the other parameters are kept fixed. Note that this is what is done in practice via the analysis-of-deviance to build additive models.

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

$$F_{AM} = \frac{\mathbf{y}^T(R_1 - R_0)\mathbf{y}}{\mathbf{y}^T(I - R_1)\mathbf{y}}, \tag{17}$$

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

The p-value is computed by

$$\begin{aligned} P(F_{AM} > f_{AM,obs}) &= P(\mathbf{y}^T(R_1 - R_0 - f_{AM,obs}(I - R_1))\mathbf{y} > 0) \\ &= P(R_{AM} > 0), \end{aligned}$$

where $R_{AM} = \mathbf{y}^T(R_1 - R_0 - f_{AM,obs}(I - R_1))\mathbf{y}$, and $f_{AM,obs}$ is the value taken by F_{AM} on the dataset.

The distribution of R_{AM} under H_0 is a linear combination of χ^2_1 variables, where the weights are the eigenvalues of the matrix $(I_n + \sum_j 1/\lambda_{j,0} Z_j Z_j^T)(R_1 - R_0 - f_{AM,obs}(I - R_1))$. This follows again from general results on the distribution of quadratic forms in normal variables.

We remark that, as in the one predictor case, the statistic (17) can be used to assess linearity at no additional cost, by testing $H_0 : \lambda_k = \lambda_{k,0} = \infty$ against $H_A : \lambda_k = \lambda_{k,1} < \infty$, and that, here again, the test statistic can be extended to test composite hypotheses of the form $H_0 : \lambda_k = \lambda_{k,0}$ versus $H_A : \lambda_k > \lambda_{k,0}$.

We illustrate the use of this procedure on the diabetes dataset (see Sockett, Daneman, Clarson, and Ehrich, 1987 and Hastie and Tibshirani, 1990). This data come 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 a set of predictors (in which we find the age of the patient and a measure of acidity) is under consideration. The data comprises $n = 43$ observations. We therefore consider the model

$$\log(\texttt{C-peptide}) = \beta_0 + f_1(\texttt{Age}) + f_2(\texttt{Base.Deficit}). \tag{18}$$

Figure 6 shows the curves fitted by smoothing splines with $\texttt{df}_{\texttt{Age}} = 2$ and $\texttt{df}_{\texttt{Base.Deficit}} = 3$. Let us focus on the variable `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`df` or 6`df`, keeping the degrees of freedom of `Base.Deficit` equal to 3. Figure 7 shows the fitted curve for $\texttt{df}_{\texttt{Age}} = 4$ (left panel) and $\texttt{df}_{\texttt{Age}} = 6$ (right panel).

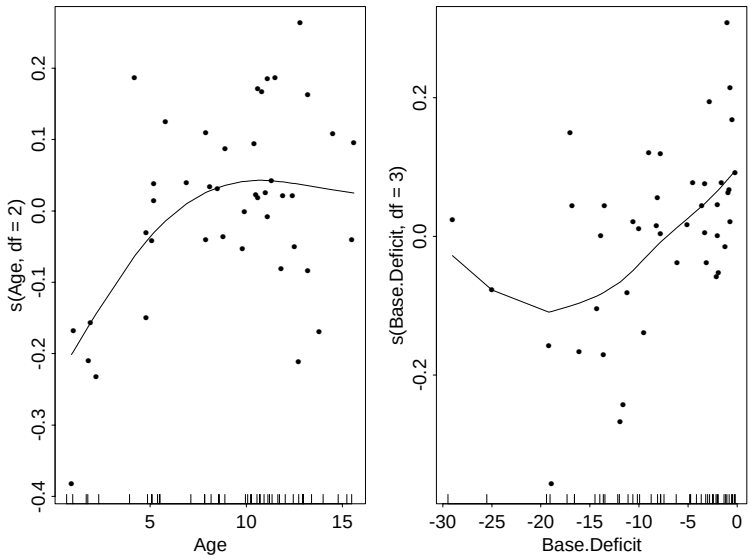

Figure 6: *Additive fit of the diabetes dataset with* $df_{\texttt{Age}} = 2$ *and* $df_{\texttt{Base.Deficit}} = 3$.

We use statistic (17) to compare these alternatives on the degrees of freedom of the variable `Age`, keeping the degrees of freedom of `Base.Deficit` fixed. The test of the null hypothesis $H_0 : \mathsf{df}_{\texttt{Age}} = 2$ versus the alternative $H_A : \mathsf{df}_{\texttt{Age}} = 4$ yields a p-value of 2×10^{-6} indicating clearly that 2 degrees of freedom are not enough.

We make use again of statistic (17) to check whether it would be beneficial to allow for more degrees of freedom. We compare the hypotheses $H_0 : \mathsf{df}_{\texttt{Age}} = 4$ and $H_A : \mathsf{df}_{\texttt{Age}} = 6$, conditional on 3 degrees of freedom for `Base.Deficit`. The p-value obtained is 0.56, which suggests that it is not worth to allow as much as 6 degrees of freedom to describe this relationship.

7 Conclusions and discussion

In this paper we propose a test statistic to choose between two predefined levels of degrees of freedom when fitting a smoothing spline. In practice, this test is used to choose the more appropriate degrees of freedom between a set of alternatives predefined by the user. Our approach is more general, but it also contains testing linearity as a particular case.

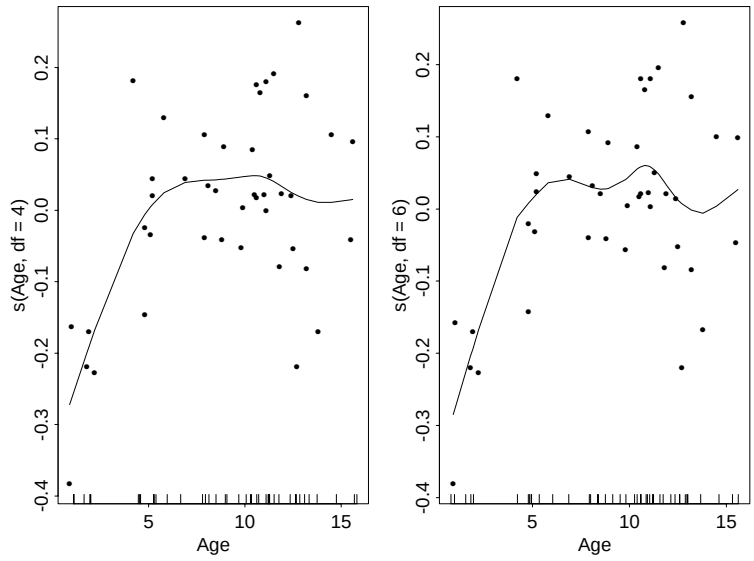

Figure 7: *More degrees of freedom for variable* `Age`*:* $df_{\texttt{Age}} = 4$ *in the left panel and* $df_{\texttt{Age}} = 6$ *in the right panel, conditional on* $df_{\texttt{Base.Deficit}} = 3$*.*

The mixed-effects model framework allows us to carry out a nice mathematical formulation and treatment of this problem in the single predictor case. The problem of bias usually encountered with nonparametric fits is avoided here thanks to this bayesian framework. Moreover, the most important feature of this procedure is the fact that the results produced are exact, at the contrary of the other procedures popualr in practice. The price to pay for this accuracy is the computational burden required to compute the parameters of the distribution of the test statistic. This is not a major concern for dataset of small or moderate sample size, but can potentially become a problem for large and huge datasets. The complexity of the algorithm is essentially due to the extraction of eigenvalues of dense matrices. Further work could improve this point, either by considering pseudo-splines (Hastie, 1996) instead of splines, or by considering numerical algorithm that intelligently extract only the largest eigenvalues. To cope with large dataset, for the moment we provide an F-approximation, which proves to works well both in terms of level and power.

The procedure is first developed for a single predictor but we successfully extended it to the multivariate case, where it provides a practical tool to

build additive models.

Moreover, our approach can be easily generalized to consider testing of composite hypotheses, both in the case of a single predictor or in additive models.

8 Acknowledgment

Eva Cantoni thanks the Swiss National Science Foundation for its support. Trevor Hastie was partially supported by grant DMS-9803645 from the National Science Foundation, and grant ROI-CA-72028-01 from the National Institutes of Health. This work has been carried out during the post-doctoral year of the first author at Stanford University.

A Appendix

A.1 Details for the mixed-effects formulation

It is well known that smoothing splines are the solution of the penalized criterion

$$J(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt. \tag{19}$$

Assuming the solution is a spline, and for a parameterization in terms of $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$, the penalty in (19) can also be written as $\mathbf{f}^T K \mathbf{f}$, with $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$, see Green and Silverman (1994) for details.

We define the eigenvalue decomposition of $K = UDU^T$ (with $UU^T = U^TU = I_n$) and partition it by block as follows

$$[U_1\ U_2] \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}, \tag{20}$$

with $D_1 = \mathrm{diag}(0,0)$ and $D_2 = \mathrm{diag}(d_3, \ldots, d_n)$ where d_i for $i = 3, \ldots, n$ are the non-zero 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 $\mathbf{x}$; in particular, we will have $U_2^T X = 0$.

Using decomposition (20), we have that the smoother matrix $S_\lambda = (I + \lambda K)^{-1}$ can be decomposed as

$$S_\lambda = U_1U_1^T + U_2(I_{n-2} + \lambda D_1)^{-1}U_2^T. \tag{21}$$

The matrix $Z = Z(\mathbf{x})$ in model (3) is defined by the relationship $U_2D_2U_2^T = (ZZ^T)^-$, see also Wahba (1990) and Speed (1991). This implies that $Z = U_2D_2^{-1/2}$, and that $Z^TX = 0$.

Moreover, the BLUP fitted values, obtained with (4) and (5), are

$$\begin{aligned}
\hat{\mathbf{y}} &= X\hat{\boldsymbol{\beta}} + Z\hat{\mathbf{u}} \\
&= X(X^TX)^{-1}X^T\mathbf{y} + Z(\lambda I_{n-2} + Z^TZ)^{-1}Z^T\mathbf{y} \\
&= \big(U_1U_1^T + U_2D_2^{-1/2}(D_2^{-1} + \lambda I_{n-2})^{-1}D_2^{-1/2}U_2^T\big)\mathbf{y} \\
&= \big(U_1U_1^T + U_2(I_{n-2} + \lambda D_2)^{-1}U_2^T\big)\mathbf{y},
\end{aligned}$$

which shows the equivalence with the fitted values obtained with the smoother matrix in (21).

A.2 Distribution of T

We have that

$$\begin{aligned}
T &= \mathbf{y}^T\big((I_n + 1/\lambda_0 ZZ^T)^{-1} - (I_n + 1/\lambda_1 ZZ^T)^{-1}\big)\mathbf{y} \\
&= (\mathbf{y} - X\boldsymbol{\beta})^T\big((I_n + 1/\lambda_0 ZZ^T)^{-1} - (I_n + 1/\lambda_1 ZZ^T)^{-1}\big)(\mathbf{y} - X\boldsymbol{\beta}) \\
&= (\mathbf{y} - X\boldsymbol{\beta})^T Z\big((\lambda_1 I_{n-2} + Z^T Z)^{-1} - (\lambda_0 I_{n-2} + Z^T Z)^{-1}\big) Z^T(\mathbf{y} - X\boldsymbol{\beta}) \\
&= (\mathbf{y} - X\boldsymbol{\beta})^T U_2\big(D_2^{-1}(\lambda_1 I_{n-2} + D_2^{-1})^{-1} - D_2^{-1}(\lambda_0 I_{n-2} + D_2^{-1})^{-1}\big)U_2^T(\mathbf{y} - X\boldsymbol{\beta}) \\
&= \mathbf{y}^{*T}\big(D_2^{-1}(\lambda_1 I_{n-2} + D_2^{-1})^{-1} - D_2^{-1}(\lambda_0 I_{n-2} + D_2^{-1})^{-1}\big)\mathbf{y}^* \\
&= \sigma^2 \mathbf{z}^T(I_{n-2} + \frac{1}{\lambda_0} D_2^{-1})\big(D_2^{-1}(\lambda_1 I_{n-2} + D_2^{-1})^{-1} - D_2^{-1}(\lambda_0 I_{n-2} + D_2^{-1})^{-1}\big)\mathbf{z} \\
&= \sigma^2 \sum_{i=3}^{n} \big(1 - \frac{d_i + 1/\lambda_0}{d_i + 1/\lambda_1}\big) z_i^2 \\
&= \sigma^2 \sum_{i=3}^{n} c_i z_i^2,
\end{aligned}$$

where $\mathbf{y}^* = U_2^T(\mathbf{y} - X\boldsymbol{\beta})$ is $\mathcal{N}(0, \sigma^2(I_{n-2} + \frac{1}{\lambda_0} D_2^{-1}))$ distributed, and $\mathbf{z} = 1/\sigma(I + \frac{1}{\lambda_0} D_2^{-1})^{-1/2}\mathbf{y}^*$ follows a standard normal distribution.

A.3 F approximation

Consider first $\sum c_i = \mathsf{Tr}((I_n + 1/\lambda_0 ZZ^T)(S_{\lambda_1} - S_{\lambda_0}))$. Having in mind that $ZZ^T = U_2 D_2^{-1} U_2^T$ and that the smoother matrix can be decomposed as $S_\lambda = U_2(I_{n-2} + \lambda D_2)^{-1}U_2^T + U_1 U_1^T$, we have that

$$\begin{aligned}
&\sum c_i = \mathsf{Tr}\big((I_n + 1/\lambda_0 ZZ^T)(S_{\lambda_1} - S_{\lambda_0})\big) = \\
&= \mathsf{Tr}\big(U_2(I_{n-2} + 1/\lambda_0 D_2^{-1})\big((I_{n-2} + \lambda_1 D_2)^{-1} - (I_{n-2} + \lambda_0 D_2)^{-1}\big)U_2^T\big) \\
&= \frac{\lambda_0 - \lambda_1}{\lambda_0}\mathsf{Tr}\big(U_2(I_{n-2} + \lambda_1 D_2)^{-1}U_2^T\big) \\
&= \frac{\lambda_0 - \lambda_1}{\lambda_0}\mathsf{Tr}\big(S_{\lambda_1} - U_1 U_1^T\big) \\
&= \frac{\lambda_0 - \lambda_1}{\lambda_0}\big(\mathsf{Tr}(S_{\lambda_1}) - 2\big) \qquad (22)
\end{aligned}$$

where we used the fact that $U_1^T U_2 = 0$.

To work out a similar formula for $\sum b_i = \mathsf{Tr}((I_n + 1/\lambda_0 ZZ^T)(I_n - S_{\tilde{\lambda}}))$, we first split it into two pieces: $\mathsf{Tr}(I_n - S_{\tilde{\lambda}}) = n - \mathsf{Tr}(S_{\tilde{\lambda}})$ and $\mathsf{Tr}(1/\lambda_0 ZZ^T(I_n - S_{\tilde{\lambda}}))$. For this second part, we have that

$$\begin{aligned} 1/\lambda_0 \mathsf{Tr}\big(ZZ^T(I_n - S_{\tilde{\lambda}})\big) &= 1/\lambda_0 \mathsf{Tr}\big(U_2 D_2^{-1}\big(I_{n-2} - (I_{n-2} + \tilde{\lambda} D_2)^{-1}\big)U_2^T\big) \\ &= \tilde{\lambda}/\lambda_0 \mathsf{Tr}\big(U_2 (I_{n-2} + \tilde{\lambda} D_2)^{-1} U_2^T\big) \\ &= \tilde{\lambda}/\lambda_0 \mathsf{Tr}\big(S_{\tilde{\lambda}} - U_1 U_1^T\big) \\ &= \tilde{\lambda}/\lambda_0 \big(\mathsf{Tr}(S_{\tilde{\lambda}}) - 2\big). \end{aligned}$$

Putting the two pieces together, we finally have that

$$\sum b_i = \mathsf{Tr}\big((I_n + 1/\lambda_0 ZZ^T)(I_n - S_{\tilde{\lambda}})\big) = n + (\tilde{\lambda}/\lambda_0 - 1)\mathsf{Tr}(S_{\tilde{\lambda}}) - 2\tilde{\lambda}/\lambda_0. \quad (23)$$

References

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

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

Chatterjee, S., Handcock, M. S., and 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 (AS R53: 84V33 p366- 369). *Applied Statistics*, **29**, 323–333.

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

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

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 379–396.

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

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *The Annals of Mathematical Statistics*, **21**, 309–310.

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*, Champaign, IL, pp. 10–41. Amer. Soc. Animal Sci. - Amer. Dairy Sci. Assn. - Poultry Sci. Assn.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*, Volume 2. Boston; Geneva, IL: Houghton-Mifflin.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models. *Journal of the Royal Statistical Society, Series B*, **61**, 381–400.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, **6**, 15–32. (Disc: p32-51).

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

Sockett, E. B., Daneman, D., Clarson, C., and Ehrich, R. M. (1987). Factors affecting and patterns of residuals insulin secretion during the first year of type I (insulin dependent) diabetes mellitus in children. *Diabetes, Journal of American Diabetes Association*, **30**, 453–459.

Speed, T. (1991). Comment on "That BLUP is a good thing: The estimation of random effects". *Statistical Science*, **6**, 42–44.

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

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, **78**, 81–89.