April, 2009

Smoothing Parameter Selection for a Class of Semiparametric Linear Models

Philip T. Reiss, *New York University*
R. Todd Ogden, *Columbia University*

Smoothing parameter selection for a class of semiparametric linear models

Philip T. Reiss[1]

New York University, New York, USA, and

Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA

and

R. Todd Ogden

Columbia University, New York, USA

September 9, 2008

[1]Address for correspondence: Philip T. Reiss, Department of Child and Adolescent Psychiatry, New York University, 215 Lexington Ave., 16th floor, New York, NY 10016.

E-mail: phil.reiss@nyumc.org

Acknowledgment. The authors thank Ciprian Crainiceanu, Eva Petkova, and Ioana Schiopu-Kratina for informative discussions; Phil Hopke for making the spectroscopy data publicly available; and the Joint Editor, Associate Editor, and referees for their comments, which have substantially improved the paper. The first author gratefully acknowledges the support of the National Institute of Mental Health through grant number 1 F31 MH73379-01A1.

Abstract

Spline-based approaches to nonparametric and semiparametric regression, as well as to regression of scalar outcomes on functional predictors, entail choosing a parameter controlling the extent to which roughness of the fitted function is penalized. In this paper we demonstrate that the equations determining two popular methods for smoothing parameter selection, generalized cross-validation and restricted maximum likelihood, share a similar form that allows us to prove several results common to both, and to derive a condition under which they yield identical values. These ideas are illustrated by application of functional principal component regression, a method for regressing scalars on functions, to two chemometric data sets.

Keywords: *B*-splines; Functional linear model; Functional principal component regression; Generalized cross-validation; Linear mixed model; Roughness penalty

1 Introduction

Roughness penalty methods are very widespread in the literature on smoothing in statistics, and especially in the spline smoothing subgenre. There is a vast body of work on fitting non- and semiparametric regression models by minimizing a loss function (such as the residual sum of squares) plus a penalty representing the roughness of the fitted function (e.g., Green and Silverman, 1994; Ruppert *et al.*, 2003). Recently there has been much interest in regressing scalar outcomes on functional predictors by similar methods (e.g., Marx and Eilers, 1999; Cardot *et al.*, 2003). Both of these classes of models fall under the rubric of the general spline problem (Wahba, 1990).

A central question in this work is how to choose the tuning parameter by which an index of roughness is multiplied to produce the roughness penalty. This parameter, often denoted by λ, is said to control the tradeoff between fidelity to the data and smoothness: too-low values of λ overfit the data, while too-high values oversmooth. Each of the two most popular approaches to smoothing parameter selection, the generalized cross-validation (GCV) method (Craven and Wahba, 1979) and the restricted maximum likelihood (REML) method (Ruppert *et al.*, 2003), proceeds by optimizing a function of λ. It seems to be common practice to tacitly assume, or hope, that the function has a unique optimum representing the ideal fidelity-smoothness tradeoff.

We undertook to investigate when such a unique optimum indeed exists. In the process we discovered that these two smoothing parameter selection methods, despite their very different motivations, can be presented within a common framework that reveals some interesting connections between them. More specifically, the derivatives of both the GCV criterion and the REML criterion with respect to λ can be expressed quite naturally in a common form. Consequently, unified arguments can serve to prove certain results concerning both criteria—including conditions under which both choose zero smoothing and maximal smoothing, as well as bounds on the

smoothing parameter. All of these results are closely related to F-tests for the smooth components of the model. The theory we develop provides insight into the sometimes surprising behavior of smoothing parameter selection methods; we illustrate this by revisiting two previously studied data sets from the field of chemometrics.

Sections 2 and 3 of this paper set out the basic model and some examples. The two smoothing parameter selection criteria are presented in Sections 4 and 5. In Section 6 we derive a one-parameter family of functions of λ such that the stationary points of both the REML and GCV criteria occur precisely where certain members of this family cross each other. This common framework is exploited in Section 7 to prove two theorems regarding the choice of λ by both criteria. Section 8 discusses our finding that the stationary points of the two criteria are the reciprocals of the positive roots of certain polynomials—raising the possibility of computing λ by solving a polynomial rather than optimizing over a grid. This section also presents a condition under which GCV and REML yield identical choices of λ. Applications of some ideas of this paper to two chemometric data sets using the functional principal component regression method of Reiss and Ogden (2007), a special case of our model, are given in Section 9. Section 10 offers some concluding remarks.

2 The general model and some special cases

We restrict our attention to linear models. The general model may be expressed in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y} \in \mathcal{R}^n$, $\boldsymbol{X}$ is $n \times p$, $\boldsymbol{Z}$ is $n \times q$, and $\boldsymbol{\varepsilon}$ is a vector of IID errors with mean 0 and variance σ^2. This model is fitted by penalized least squares, i.e., our estimate is

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}) = \underset{\boldsymbol{\beta},\boldsymbol{u}}{\arg\min} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}\|^2 + \lambda \boldsymbol{u}^T \boldsymbol{W} \boldsymbol{u} \right) \tag{2}$$

where the $q \times q$ matrix $\boldsymbol{W}$ is chosen so that $\boldsymbol{u}^T\boldsymbol{W}\boldsymbol{u}$ represents the roughness of the fitted function.

One example is Eilers' (1999) P-spline-based nonparametric regression model

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{3}$$

Let $\boldsymbol{B} = [b_j(t_i)]_{1 \le i \le n, 1 \le j \le q}$ where $b_1, \ldots, b_q$ form a cubic B-spline basis. In Eilers' formulation (slightly adapted to our notation), $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X}^*$ is an $n \times 2$ matrix whose columns are a constant and a linear function, and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}^T(\boldsymbol{D}\boldsymbol{D}^T)^{-1}$, where $\boldsymbol{D}$ is a $(q-2) \times q$ second-order differencing matrix. As indicated by the inclusion of $\boldsymbol{u}$, but not $\boldsymbol{\beta}$, in the penalty in (2), the columns of $\boldsymbol{X}$ represent functions of zero roughness.

Equation (1) likewise applies when we pass from the nonparametric regression model (3) to the semiparametric regression model

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + f(t_i) + \varepsilon_i, \tag{4}$$

where $\boldsymbol{x}_i$ is a vector of covariates. For this model, sometimes referred to as a semilinear or partial spline model, a column corresponding to each covariate is added to $\boldsymbol{X}$.

Another special case of (1) is the functional linear model

$$y_i = \beta + \int s_i(t) f(t) dt + \varepsilon_i, \quad i = 1, \ldots, n,$$

where s_i is a signal, or functional predictor, for the ith subject, and f is the coefficient function we seek to estimate. If the signals are discretized at a grid of points $t_1, \ldots, t_N$, and f is restricted to the span of spline functions $b_1, \ldots, b_q$, then this model can be written as (1), with $\boldsymbol{X}$ a column of ones and $\boldsymbol{Z} = \boldsymbol{S}\boldsymbol{B}$, where $\boldsymbol{S} = [s_i(t_j)]_{1 \le i \le n, 1 \le j \le N}$ and $\boldsymbol{B}$ is now the $N \times q$ matrix $[b_j(t_i)]_{1 \le i \le N, 1 \le j \le q}$. The model is fitted by minimizing

$$\|\boldsymbol{y} - \beta\boldsymbol{1} - \boldsymbol{S}\boldsymbol{B}\boldsymbol{u}\|^2 + \lambda \boldsymbol{u}^T\boldsymbol{W}\boldsymbol{u}, \tag{5}$$

where $\boldsymbol{u}^T\boldsymbol{W}\boldsymbol{u}$ is an index of the roughness of the coefficient function given in discretized form by $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{u}$. Marx and Eilers (1999) take $\boldsymbol{W} = \boldsymbol{D}^T\boldsymbol{D}$ where $\boldsymbol{D}$ is a

kth-order differencing matrix. This roughness penalty approximates the integrated squared kth derivative penalty used by other authors such as Cardot *et al.* (2003). As above, we can extend this functional linear model to a semiparametric model by adding covariates (cf. the more general model of Eilers and Marx (2002)).

In the nonparametric regression case, λ is often chosen either by GCV or by REML (see below, Sections 4 and 5). In the functional regression literature, Marx and Eilers (1999) proposed ordinary cross-validation, while Cardot *et al.* (2003) used GCV; REML yielded superior results in the simulations of Reiss and Ogden (2007).

3 Semiparametric model assumptions

The development that follows depends on the following assumptions:

1. $\boldsymbol{X}$ is a full-rank $n \times p$ matrix, and $\boldsymbol{Z}$ a full-rank $n \times q$ matrix, with $p + q \leq n$.
2. $\boldsymbol{W}$ is invertible.
3. $\boldsymbol{X}^T \boldsymbol{Z} = \boldsymbol{0}$.
4. $\boldsymbol{y}$ is not in the column space of $\boldsymbol{X}$.

Assumptions 1 and 4 are mild technical conditions; Assumptions 2 and 3 are more restrictive. When (1) represents the nonparametric regression model (3) or the semiparametric regression model (4), Assumption 2 may not hold, and Assumption 3 does not hold in many formulations (although Cantoni and Hastie (2002) do make this assumption). However, some matrix manipulations lead to a reparametrized model for which both assumptions hold. We discuss next the meaning of Assumptions 2 and 3 for the functional linear model minimizing (5), our primary model of interest here.

In view of criterion (5), Assumption 2 means that any nonzero coefficient function $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{u}$ has a positive penalty. This is generally false for a difference penalty

$\boldsymbol{W} = \boldsymbol{D}^T\boldsymbol{D}$, but it is true for a derivative penalty in some settings. Assumption 2 is not at all restrictive if we stipulate $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{u}^*$ for a dimension-reducing matrix $\boldsymbol{V}$ and choose $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}^*)$ to minimize $\|\boldsymbol{y} - \beta\mathbf{1} - \boldsymbol{S}\boldsymbol{B}\boldsymbol{V}\boldsymbol{u}^*\|^2 + \lambda\boldsymbol{u}^{*T}\boldsymbol{W}\boldsymbol{u}^*$ with $\boldsymbol{W} = \boldsymbol{V}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}$. For instance, Reiss and Ogden (2007) obtained convergence theorems, and favourable empirical results with real and simulated data, by taking $\boldsymbol{V} = (\boldsymbol{v}_1 \dots \boldsymbol{v}_q)$ such that $\boldsymbol{S}\boldsymbol{B}\boldsymbol{V}$ gives the first q principal components, or first q partial least squares components, of $\boldsymbol{S}\boldsymbol{B}$. They referred to these methods as *functional principal component regression* (FPCR) and *functional partial least squares* (FPLS), respectively. The former method is used below to illustrate some of our results.

Assumption 3, in the case of functional regression with no covariates, reduces to $\mathbf{1}^T\boldsymbol{Z} = \mathbf{0}$. Thus if $\boldsymbol{Z} = \boldsymbol{S}\boldsymbol{M}$ for some matrix $\boldsymbol{M}$, as is the case for FPCR/FPLS, the assumption holds provided $\mathbf{1}^T\boldsymbol{S} = \mathbf{0}$, as is indeed commonly imposed in signal regression by subtracting out the mean from each column. If covariates are added, Assumption 3 will still hold if we "decorrelate" the signals from the covariates, i.e., replace each column of $\boldsymbol{S}$ with the residuals from a regression of that column on $\boldsymbol{X}$.

4 Choosing λ by restricted maximum likelihood

As our choice of matrix notation suggests, model (1) can be seen as equivalent to a linear mixed model, in the following sense. The criterion in (2) is proportional to the log likelihood for the partly observed "data" $(\boldsymbol{y}, \boldsymbol{u})$ with respect to the unknowns $\boldsymbol{\beta}$ and $\boldsymbol{u}$, i.e., the best linear unbiased prediction (BLUP) criterion, for the mixed model

$$\boldsymbol{y}|\boldsymbol{u} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2\boldsymbol{I}),\ \boldsymbol{u} \sim N[0, (\sigma^2/\lambda)\boldsymbol{W}^{-1}],$$

where we have used Assumption 2 to ensure the existence of $\boldsymbol{W}^{-1}$. Under this model, $\mathrm{Var}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}_\lambda$ where

$$\boldsymbol{V}_\lambda = \boldsymbol{I} + \lambda^{-1}\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T. \tag{6}$$

The mixed model formulation motivates treating λ as a variance parameter to be estimated by maximizing the log likelihood

$$l(\boldsymbol{\beta}, \lambda, \sigma|\boldsymbol{y}) = -\frac{1}{2}\left[\log|\sigma^2\boldsymbol{V}_\lambda| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\boldsymbol{V}_\lambda)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right].$$

Maximizing this log likelihood results in estimating σ^2 with a downward bias, which is removed if we instead maximize the restricted log likelihood

$$l_R(\boldsymbol{\beta}, \lambda, \sigma|\boldsymbol{y}) = -\frac{1}{2}\left[\log|\sigma^2\boldsymbol{V}_\lambda| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\sigma^2\boldsymbol{V}_\lambda)^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + \log|\sigma^{-2}\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|\right]. \tag{7}$$

We shall refer to the resulting estimate of λ as the REML choice of the parameter, although some authors favour the term "generalized maximum likelihood."

Under our assumptions, the matrix

$$\boldsymbol{P}_\lambda = \boldsymbol{V}_\lambda^{-1} - \boldsymbol{V}_\lambda^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1},$$

which plays a role in some treatments of mixed model theory, turns out to be important for both the REML and the GCV approach to choosing λ. Appendix A shows that $l_R(\lambda)$, the profile restricted log likelihood with respect to λ alone, has derivative

$$\frac{dl_R(\lambda|\boldsymbol{y})}{d\lambda} = \frac{1}{2\lambda}\left[(n-p)\frac{(\boldsymbol{y}^T\boldsymbol{P}_\lambda^2\boldsymbol{y})}{(\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y})} - \mathrm{tr}(\boldsymbol{P}_\lambda)\right]. \tag{8}$$

Thus by (24), (28) and (8), $\frac{d}{d\lambda}l_R(\lambda|\boldsymbol{y}) = 0$ implies

$$\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)^T\boldsymbol{V}_\lambda^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)}{n-p} = \frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{H}_\lambda)^2\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{I}-\boldsymbol{H}_\lambda)}, \tag{9}$$

where $\hat{\boldsymbol{\beta}}_\lambda$ and $\boldsymbol{H}_\lambda$ are the parameter estimate and hat matrix, respectively, obtained with smoothing parameter value λ (see Appendix A for explicit expressions). The left side of (9) is the REML estimate of σ^2 [cf. (22)]. The right side equals $\|\boldsymbol{y}-\hat{\boldsymbol{y}}\|^2/[n-\mathrm{tr}(\boldsymbol{H}_\lambda)]$, an estimate of σ^2 based on viewing $\mathrm{tr}(\boldsymbol{H}_\lambda)$ as the degrees of freedom of the smoother (cf. Pawitan, 2001, p. 487, and Lee, Nelder, and Pawitan, 2006, p. 279). In other words, when λ is estimated by REML, the REML error variance estimate agrees with the "smoothing-theoretic" variance estimate.

5 Choosing λ by GCV

The GCV criterion is given by

$$GCV(\lambda) = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2/n}{[1 - \mathrm{tr}(\boldsymbol{H}_\lambda)/n]^2} = \frac{n\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_\lambda)^2\boldsymbol{y}}{[\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}_\lambda)]^2} = \frac{n\boldsymbol{y}^T\boldsymbol{P}_\lambda^2\boldsymbol{y}}{[\mathrm{tr}(\boldsymbol{P}_\lambda)]^2},$$

with the last equality following from (24). This criterion, originally proposed by Craven and Wahba (1979), is an approximation to $\frac{1}{n}\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{(1-h_{\lambda ii})^2}$, where $h_{\lambda 11}, \ldots, h_{\lambda nn}$ are the diagonal elements of $\boldsymbol{H}_\lambda$. The latter expression can be shown (at least in some smoothing problems) to be equal to the leave-one-out cross-validation criterion, but lacks an invariance-under-reparametrization property that is gained by instead using GCV (Wahba, 1990, pp. 52–53). Using (31), we can obtain

$$\frac{d}{d\lambda}GCV(\lambda) = \frac{2n}{\lambda[\mathrm{tr}(\boldsymbol{P}_\lambda)]^3}\left[\mathrm{tr}(\boldsymbol{P}_\lambda^2)\boldsymbol{y}^T\boldsymbol{P}_\lambda^2\boldsymbol{y} - \mathrm{tr}(\boldsymbol{P}_\lambda)\boldsymbol{y}^T\boldsymbol{P}_\lambda^3\boldsymbol{y}\right]. \tag{10}$$

Thus at the GCV-minimizing λ we have

$$\frac{\boldsymbol{y}^T\boldsymbol{P}_\lambda^3\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{P}_\lambda^2)} = \frac{\boldsymbol{y}^T\boldsymbol{P}_\lambda^2\boldsymbol{y}}{\mathrm{tr}(\boldsymbol{P}_\lambda)}.$$

6 Stationary points of the REML and GCV functions: a common framework

As alluded to in Section 1, we ordinarily expect $l_R(\lambda|\mathbf{y})$ to be an increasing function of λ up to a unique maximum $\hat{\lambda}_{REML}$, and a decreasing function thereafter; we would likewise expect $GCV(\lambda)$ to be decreasing until a unique minimum $\hat{\lambda}_{GCV}$ and increasing thereafter. For brevity let us call this the "well-behaved" situation. The GCV and REML criteria may fail to behave well in three ways:

(I) the GCV (REML) criterion attains its lowest (highest) point at $\lambda = 0$;

(II) the GCV (REML) criterion has no minimum (maximum), but becomes progressively smaller (larger) as $\lambda \to \infty$, so that the criterion chooses $\lambda = \infty$;

(III) GCV (REML) has multiple minima (maxima) on $(0, \infty)$.

Some understanding of when these various situations obtain can be gained by studying the stationary points of the restricted likelihood and GCV, viewed as functions of λ. As we show in this section, stationary points of both functions are equivalent to mutual crossings of certain members of a one-parameter class of functions. This makes possible a theoretical development that encompasses both REML and GCV.

Our starting point is the singular value decomposition $\boldsymbol{P}_\lambda = \sum_{i=1}^n d_{\lambda i} \boldsymbol{u}_i \boldsymbol{u}_i^T$. It will simplify the notation if we depart from the convention of indexing the eigenvectors in descending order, and divide the eigenvectors $\boldsymbol{u}_i$ and associated eigenvalues $d_{\lambda i}$ into three groups, as follows (see Appendix B for details):

1. For $i = 1, \ldots, q$, eigenvector $\boldsymbol{u}_i$ corresponds to eigenvalue $d_{\lambda i} = \frac{\lambda}{\lambda + \gamma_i}$, where $\gamma_1 \geq \ldots \geq \gamma_q$ are the positive eigenvalues of $\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T$.

2. Eigenvectors $\boldsymbol{u}_{q+1}, \ldots, \boldsymbol{u}_{n-p}$ correspond to eigenvalue $d_{\lambda,q+1} = \ldots d_{\lambda,n-p} = 1$.

3. Eigenvectors $\boldsymbol{u}_{n-p+1}, \ldots, \boldsymbol{u}_n$ correspond to eigenvalue $d_{\lambda,n-p+1} = \ldots = d_{\lambda n} = 0$.

Let $y_{(1)}, \ldots, y_{(n)}$ be the coordinates of $\boldsymbol{y}$ with respect to the basis formed by the $\boldsymbol{u}_i$'s, i.e., $y_{(i)} = \boldsymbol{y}^T \boldsymbol{u}_i$ for $i = 1, \ldots, n$. For $k = 1, 2, \ldots$, define

$$\begin{aligned} Q_{\lambda k} &= \boldsymbol{y}^T \boldsymbol{P}_\lambda^k \boldsymbol{y} \\ &= \sum_{i=1}^n y_{(i)}^2 d_{\lambda i}^k \\ &= \sum_{i=1}^q y_{(i)}^2 \left(\frac{\lambda}{\lambda + \gamma_i} \right)^k + \sum_{i=q+1}^{n-p} y_{(i)}^2 \end{aligned} \tag{11}$$

and

$$t_{\lambda k} = \mathrm{tr}(\boldsymbol{P}_\lambda^k) = n - p - q + \sum_{i=1}^q \left(\frac{\lambda}{\lambda + \gamma_i} \right)^k . \tag{12}$$

In light of the latter expression for $t_{\lambda k}$, we can also reasonably define $t_{\lambda 0} = n - p$. Finally, for $k = 1, 2, \ldots$, define $h_k(\lambda) = Q_{\lambda k} / t_{\lambda, k-1}$. It follows from (8), (10) that

$$\mathrm{sgn} \frac{dl_R(\lambda | \boldsymbol{y})}{d\lambda} = \mathrm{sgn}[h_2(\lambda) - h_1(\lambda)] \text{ and} \tag{13}$$

$$\operatorname{sgn}\left[\frac{d}{d\lambda}GCV(\lambda)\right] = \operatorname{sgn}[h_2(\lambda) - h_3(\lambda)]. \tag{14}$$

Equations (13) and (14) enable us to characterize the well-behaved situation in terms of the functions h_1, h_2, and h_3. In this situation we have

$$h_1(\lambda) < h_2(\lambda) < h_3(\lambda) \tag{15}$$

for sufficiently small $\lambda > 0$, while

$$h_1(\lambda) > h_2(\lambda) > h_3(\lambda) \tag{16}$$

for λ sufficiently large. Moreover, in the well-behaved situation, h_1 crosses h_2 (from smaller to larger) at a unique value $\hat{\lambda}_{REML}$, whereas h_3 crosses h_2 (from larger to smaller) at a unique point $\hat{\lambda}_{GCV}$. REML smooths more than GCV if and only if the latter crossing of h_2 precedes the former one, as in Figure 1.

From the definitions above we see that, as λ goes from 0 to ∞, $Q_{\lambda k}$ $(k = 1, 2, 3)$ increases from $\sum_{i=q+1}^{n-p} y_{(i)}^2$ to $\sum_{i=1}^{n-p} y_{(i)}^2$; $t_{\lambda 1}$ and $t_{\lambda 2}$ increase from $n - p - q$ to $n - p$; but $t_{\lambda 0}$ remains constant at $n - p$. Hence, as λ increases, the numerators and denominators of h_2 and h_3 both increase, but h_1 has an increasing numerator and a constant denominator. This explains why, in Figure 1, h_1 rises more sharply than the other two functions, and thus h_1 crosses h_2 at a sharper angle than does h_3. We have observed qualitatively similar behavior with different data sets, different bases, and different numbers of principal components. This heuristic argument suggests that $\hat{\lambda}_{REML}$ will tend to be more stable than $\hat{\lambda}_{GCV}$. One manifestation of this is that $\hat{\lambda}_{REML}$ generally seems to increase with the number of components (as might be expected, in order to maintain approximately the same degrees of freedom), whereas $\hat{\lambda}_{GCV}$ seems much less predictable in this regard. In agreement with these observations, Kauermann (2005), employing somewhat different assumptions, showed $\hat{\lambda}_{REML}$ to have lower variance than the smoothing parameter chosen by the C_p criterion, which is asymptotically equivalent to GCV (cf. Kou, 2004).

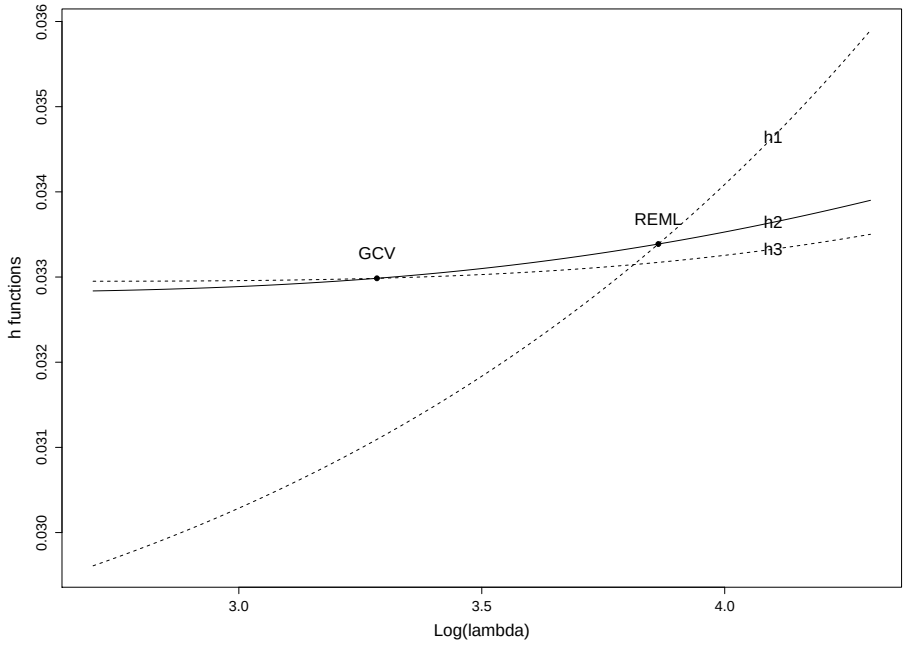

Figure 1: Graphical representation of the choice of λ for an FPCR fit. The functions h_1, h_2, and h_3 are shown for a 10-component FPCR model using the gasoline data described below in Section 9.3. The function h_3 (dotted line) crosses below h_2 (solid line) at $\log(\hat{\lambda}_{GCV})$; h_1 (dashed line) crosses above h_2 at $\log(\hat{\lambda}_{REML})$. Since $\hat{\lambda}_{GCV} < \hat{\lambda}_{REML}$, REML smooths more than GCV in this instance.

7 When do REML and GCV choose $\lambda \in (0, \infty)$?

This section and Section 8 present conditions under which one or more of Scenarios (I)-(III) either hold, or can be ruled out, for both GCV and REML simultaneously. These results can be derived from statements about h_1, h_2, and h_3: for instance, if (15) holds for all $\lambda > 0$ then (II) obtains, in both its GCV and its REML version.

For $i = 1, \ldots, q$, let $F_{(i)} = \frac{y_{(i)}^2}{\sum_{i=q+1}^{n-p} y_{(i)}^2/(n-p-q)}$. This is the F-statistic for removing $\boldsymbol{u}_i$ from the ordinary linear model with design matrix $(\boldsymbol{X}\ \boldsymbol{u}_1 \ldots \boldsymbol{u}_q)$, which has the same column space as $(\boldsymbol{X}\ \boldsymbol{Z})$. F-statistics of this type have little or no practical use since $\boldsymbol{u}_i$ is not a column of the actual design matrix, but they play a role in our development because, for $m = 1, 2, 3$, $h_m(\lambda)(n-p-q)/\sum_{i=q+1}^{n-p} y_{(i)}^2$ is equal to—and hence $h_m(\lambda)$ is proportional to—the weighted average of

$$F_{(1)} d_{\lambda 1}, \ldots, F_{(q)} d_{\lambda q}, 1 \tag{17}$$

with weights $d_{\lambda 1}^{m-1}, \ldots, d_{\lambda q}^{m-1}, n-p-q$. Since $1 > d_{\lambda i} > d_{\lambda i}^2$ for all $\lambda > 0$ and for $i = 1, \ldots, q$, we see that the first q quantities in (17) are accorded less weight in $h_2(\lambda)$ than in $h_1(\lambda)$, but more weight in $h_2(\lambda)$ than in $h_3(\lambda)$. It is intuitively clear, then, that at values of λ at which h_2 crosses either h_1 or h_3, the first q quantities in (17) are neither all greater, nor all smaller, than the $(q+1)$th. This intuitive assertion is given a more formal and more general statement as Lemma 1 in Appendix C.

Observe that $\sum_{i=q+1}^{n-p} y_{(i)}^2$ is the squared norm of the projection of $\boldsymbol{y}$ onto the orthogonal complement of the combined column space of $\boldsymbol{X}$ and $\boldsymbol{Z}$, and so this expression equals 0 if and only if $\boldsymbol{y}$ can be recovered without error as a linear combination of the columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$. Our first result says that each of the two criteria chooses $\lambda = 0$ precisely in this very special case. Moreover, in the contrary case, we can establish bounds on λ depending on which of the statements $F_{(i)} \leq 1$, $i = 1, \ldots, q$, hold.

Theorem 1 *Under Assumptions 1–4,*

(i) if $\sum_{i=q+1}^{n-p} y_{(i)}^2 = 0$ *then both GCV and REML choose* $\lambda = 0$*;*

(ii) if $\sum_{i=q+1}^{n-p} y_{(i)}^2 > 0$ *then both GCV and REML choose a positive* λ *or* $\lambda = \infty$*. More specifically, for both criteria,*

(a) if $F_{(i)} \leq 1$ *for* $i = 1, \ldots, q$ *then the unique optimum is* $\lambda = \infty$*;*

(b) if $F_{(i)} > 1$ *for some* $i \in \{1, \ldots, q\}$*, then all local optima are in*

$$\left[\min_{\{1 \leq i \leq q,\ F_{(i)} > 1\}} \frac{\gamma_i}{F_{(i)} - 1},\ \infty \right];$$

(c) if $F_{(i)} > 1$ *for all* $i \in \{1, \ldots, q\}$*, then all local optima are in*

$$\left[\min_{1 \leq i \leq q} \frac{\gamma_i}{F_{(i)} - 1},\ \max_{1 \leq i \leq q} \frac{\gamma_i}{F_{(i)} - 1} \right].$$

Roughly speaking, Theorem 1(ii) tells us that when $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$ are quite (positively or negatively) correlated with $\boldsymbol{y}$, λ will tend to be small. This makes intuitive sense since the former condition indicates that $\boldsymbol{Z}$ predicts $\boldsymbol{y}$ successfully and thus relatively little smoothing is needed. A similar intuitive explanation applies to our next result, which provides a weaker sufficient condition for finite λ than that of Theorem 1(ii)(c) (albeit without providing an upper bound for λ, as that result does).

Theorem 2 *Under Assumptions 1–4,*

(i) if

$$\frac{\sum_{i=1}^{q} \gamma_i y_{(i)}^2}{\sum_{i=1}^{q} \gamma_i} < \frac{\sum_{i=1}^{n-p} y_{(i)}^2}{n-p} \tag{18}$$

then GCV and REML each have at least a local optimum $\lambda = \infty$ *(i.e., GCV is decreasing, and the restricted likelihood increasing, for all sufficiently large* λ*);*

(ii) if the opposite strict inequality holds then both criteria choose $\lambda < \infty$ *.*

When the left and right sides of (18) are equal, the criteria may or may not have a local optimum at ∞, although our experience with simulated data suggests that they usually do. This theorem can, like the previous one, be expressed in terms of F-statistics, by noting that (18) is equivalent to

$$\frac{\sum_{i=1}^{q} \gamma_i (F_{(i)} - 1)}{\sum_{i=1}^{q} \gamma_i} < \frac{\sum_{i=1}^{q} (F_{(i)} - 1)}{n - p}.$$

Theorem 2 unifies several existing results on testing the null hypothesis $\lambda = \infty$, which have appeared separately in discussions of GCV and of REML. Since the non-parametric component in (1) becomes negligible as $\lambda \to \infty$, testing this null can be thought of as testing the null model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{19}$$

versus the alternative (1). Specifically, with some translation of notation, Theorem 2(i) yields a sufficient condition for GCV to have a possibly local minimum at $\lambda = \infty$ (Theorem 3 of Cox *et al.*, 1988, and Theorem 6.3.1 of Wahba, 1990). Likewise, since $y_{(i)}^2 \sim \sigma^2 \chi_1^2$ for $i = 1, \ldots, n - p$ under the null model (19), Theorem 2 can be used to re-derive the expression of Crainiceanu and Ruppert (2004, p. 170) for the probability of a local maximum of the REML criterion at $\lambda = \infty$.

8 Local optima at the reciprocals of the positive zeros of a polynomial

In Appendix D, using the quantities $Q_{\lambda,k}$ and $t_{\lambda,k}$ defined in (11) and (12), we derive a $(2q-1)$-degree polynomial in $\theta = 1/\lambda$ such that the reciprocals of the positive zeros of the polynomial are the stationary points of $l_R(\lambda)$. A similar expression defines a $(3q - 2)$-degree polynomial whose positive zeros correspond in the same way to

the stationary points of $GCV(\lambda)$. This finding indicates that l_R may have multiple maxima, and GCV multiple minima. The latter phenomenon was connected by Hastie and Tibshirani (1990) with the tendency of GCV to undersmooth in some cases. A study of the number of positive zeros of these polynomials lies beyond the scope of this paper, but the higher degree of the polynomial for GCV than for l_R seems to indicate a higher probability of multiple positive zeros. We have begun to explore the use of a symbolic computing package to compute the coefficients of the polynomial and obtain its positive roots. Consequently, one would need only to check the value of the GCV or REML criterion at the reciprocal of these roots, rather than searching over a range of values, to obtain the global optimum.

In the special case in which all of the positive eigenvalues of $\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T$ are equal, the polynomial of Appendix D simplifies radically, and surprisingly, the GCV and REML criteria then become equivalent. This is stated formally in the result below, which is proved in Appendix C. Let

$$F_{(\mathbf{Z})} = \sum_{i=1}^{q} F_{(i)}/q = \frac{\sum_{i=1}^{q} y_{(i)}^2/q}{\sum_{i=q+1}^{n-p} y_{(i)}^2/(n-p-q)}, \tag{20}$$

the F-statistic for comparing the linear models with design matrices $\boldsymbol{X}$ and $(\boldsymbol{X}\ \boldsymbol{Z})$, i.e., testing the null model (19) against the alternative (1).

Theorem 3 *Suppose that Assumptions 1–4 hold,* $\sum_{i=q+1}^{n-p} y_{(i)}^2 > 0$, *and* $\gamma_1 = \ldots = \gamma_q = \gamma$. *Then the GCV and REML criteria each attain a unique optimum*

(i) at $\lambda = \frac{\gamma}{F_{(\mathbf{Z})}-1}$, *if* $F_{(\mathbf{Z})} > 1$; *and*

(ii) at $\lambda = \infty$, *otherwise.*

9 Illustrations with real and simulated data

9.1 Data sets

We illustrate some of the ideas of the preceding sections using two data sets described in Kalivas (1997) and publicly available at Prof. Phil Hopke's ftp site, ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/. The first data set consists of near-infrared reflectance spectra of 100 wheat samples, measured in 2-nm intervals from 1100 to 2500 nm, and an associated response variable, the samples' moisture content. The ability to predict moisture in a wheat sample by spectroscopic methods has great practical value because high moisture content can lead to storage problems for wheat. To correct for a baseline shift observed in the wheat spectra, we used the once-differenced spectra. The second data set comprises spectra from 60 gasoline samples, measured in 2-nm intervals from 900 to 1700 nm, along with the samples' octane numbers.

9.2 Effect of number of components on the choice of λ

Figure 2(a)–(c) displays $\boldsymbol{B}\boldsymbol{v}_1, \boldsymbol{B}\boldsymbol{v}_2, \boldsymbol{B}\boldsymbol{v}_3$ for the wheat spectra, where $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ are the first three principal components of the matrix $\boldsymbol{S}\boldsymbol{B}$ appearing in (5). Here $\boldsymbol{S}$ is the 100×700 matrix of spectra and $\boldsymbol{B}$ is a 700×43 matrix whose columns are discretized versions of B-splines. Choosing the number of components by 5-fold cross-validation resulted in a 10-component FPCR model for predicting moisture content, whether using GCV or REML to select λ. The GCV-based 10-component coefficient function estimate is represented by the solid curve in Figure 2(d). If we assume this estimate is fairly close to the true coefficient function, we can infer that the main feature of the latter is a trough near 2035 nm. Since the first two components are quite featureless around that wavelength, while the third has its main peak nearby, we would expect

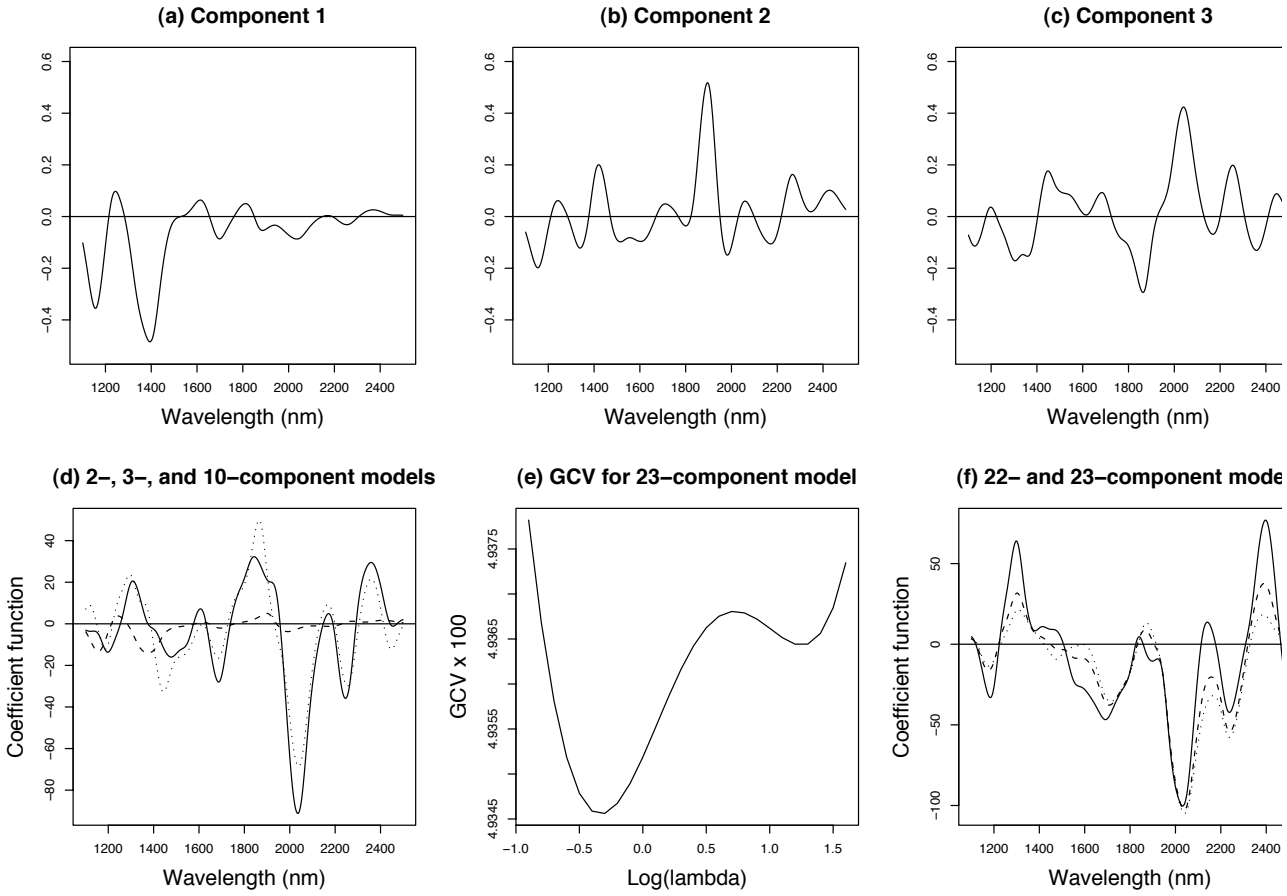

Figure 2: Illustration of FPCR with varying numbers of principal components. The main peak of the third component (c) is near the main feature of the 10-component coefficient function estimate (solid line in (d)), the estimate chosen as optimal by cross-validation. This enables the 3-component estimate (dotted line in (d)) to approach the optimal estimate much more closely than does the 2-component estimate (dashed line). Subfigure (e) shows the occurrence of two minima of the GCV function for the 23-component model. The minimum at right yields a fit (dashed line in subfigure (f)) quite similar to the 22-component fit (dotted line), while the minimum at left leads to a quite different fit (solid line).

a 3-component model to be more successful than a 2-component model. And indeed the 3-component estimate (dotted curve in Figure 2(d)) does share most features of the 10-component estimate (solid curve), whereas the 2-component estimate (dashed curve) does not. Interestingly, as one passes from the 2- to the 3-component model, the range of possible values for λ, based on Theorem 1(ii)(c), changes from [253, 731] to [.30, 9.98]. This change reflects the fact that no linear combination of the first two components predicts the moisture content satisfactorily, and consequently the best strategy is to "cut our losses" with a very smooth fit (large λ); whereas the third component greatly improves our ability to predict the outcome, and this obviates the need for such extreme smoothing.

Figure 2(e) displays GCV as a function of λ for the 23-component model. GCV has a global minimum at .72 as well as a local minimum at 3.49. Figure 2(f) compares the 23-component coefficient function estimates obtained with these two values of λ. The $\lambda = 3.49$ fit (dashed curve) is more similar to the 22-component fit (dotted curve) than to the $\lambda = .72$ fit (solid curve). Evidently, the minimum at .72 represents a transition to a markedly different solution upon addition of the 23rd component. We remark that the REML value $\lambda = 3.11$ for the 23-component model is much closer to the second minimum than to the global minimum of the GCV function.

9.3 Multiple minima of GCV in a simulation study

The phenomenon of multiple minima of GCV as a function of λ was explored in a small simulation study using the gasoline signals. We generated 300 sets of continuous outcomes by taking inner products of the spectra with the bumpy function (true coefficient function) used by Reiss and Ogden (2007), and adding Gaussian noise such that the coefficient of determination for the true model was 0.9. For each set of outcomes, the coefficient function was estimated with 30-component FPCR, with

λ chosen by GCV and by REML. Let λ_{gm} denote the value of λ at which $GCV(\lambda)$ attains its global minimum, and let λ_* denote the largest λ at which $GCV(\lambda)$ attains a local minimum. In 14 of the 300 simulations, we found that $\lambda_{gm} < \lambda_*$. (The corresponding phenomenon, with local and global maxima rather than minima, did not occur for REML in any of the 300 simulations.) In each of these 14 cases, λ_{gm} led to substantial overfitting as gauged by degrees of freedom of the fit, $df(\lambda) \equiv \mathrm{tr}(\boldsymbol{H}_\lambda)$. Using λ_* instead of λ_{gm} results in lower degrees of freedom, as shown in Figure 3, but the df remained higher than the value that is optimal in the sense of minimizing the coefficient function estimate's mean integrated squared error (Reiss, 2006; Reiss and Ogden, 2007). In other words, in each case, using λ_* instead of λ_{gm} remedied the overfitting without underfitting.

9.4 Effect of adjusting the noise level

Another illustration of multiple minima of GCV is provided by a set of FPCR models we fitted with perturbed versions of the outcome $\boldsymbol{y}$ (octane number) accompanying the gasoline spectra. To define perturbed outcomes, we posited that the statistic $F_{(\mathbf{Z})}$ in (20) is distributed as $F_{q,n-p-q,\Delta}$, whose noncentrality parameter Δ we estimated by setting $F_{(\mathbf{Z})} = E(F_{q,n-p-q,\hat{\Delta}})$. The actual outcome vector has a unique decomposition $\boldsymbol{y} = \boldsymbol{P}_{(0)}\boldsymbol{y} + \boldsymbol{P}_{(1)}\boldsymbol{y}$, where $\boldsymbol{P}_{(0)}$, $\boldsymbol{P}_{(1)}$ denote projection onto the column space of $(\boldsymbol{X}\ \boldsymbol{Z})$ and onto its orthogonal complement, respectively. Our perturbed outcomes were given by $\boldsymbol{y}^*(c) = \boldsymbol{P}_{(0)}\boldsymbol{y} + c\boldsymbol{P}_{(1)}\boldsymbol{y}$, with c chosen such that the resulting statistic $F^*_{(\mathbf{Z})}(c)$ computed in analogy to (20) would equal a range of quantiles of the $F_{q,n-p-q,\hat{\Delta}}$ distribution. We then fitted FPCR models using REML and GCV to obtain $\lambda_1^*(c)$ and $\lambda_2^*(c)$, respectively. If we imagine c increasing through the range $(0, \infty)$, $F^*_{(\mathbf{Z})}(c)$ will decrease throughout that same range, whereas, by Theorem 1(ii), both $\lambda_1^*(c)$ and

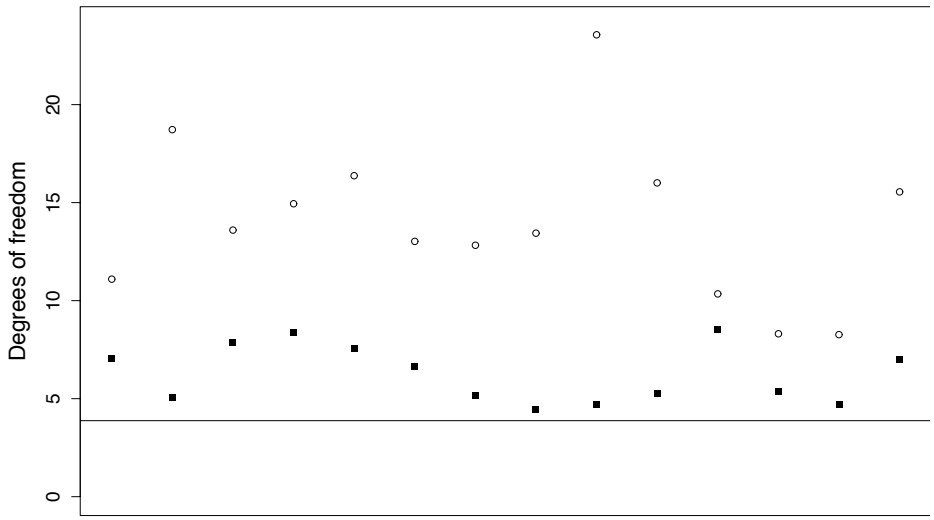

Figure 3: Degrees of freedom of the fits derived from two GCV-based choices of λ. The circles indicate degrees of freedom of the fits obtained with $\lambda = \lambda_{gm}$, for the 14 simulations in which $\lambda_{gm} < \lambda_*$. The black square below each circle indicates the degrees of freedom resulting when we instead take $\lambda = \lambda_*$. In each case, the latter choice brings the df closer to the optimal df indicated by the horizontal line.

$\lambda_2^*(c)$ will increase, beginning near 0 and eventually attaining ∞; hence both

$$P[F_{q,n-p-q,\hat{\Delta}} \leq F^*_{(\mathbf{Z})}(c)] \tag{21}$$

and $g[\lambda_k^*(c)] \equiv \frac{1}{1+\lambda_k^*(c)/\lambda_2}$ will essentially pass from 0 to 1, for $k = 1, 2$, where $\lambda_2 \equiv \lambda_2^*(1)$ is the GCV value obtained with the actual outcome.

Plots of cumulative density (21) versus $g[\lambda_k^*(c)]$ are shown for 10-, 20-, 30-, and 40-component models in Figure 4. They are generally reasonably flat, indicating insensitivity of the choice of λ to perturbations of this kind. However, with 30 and 40 components, the GCV curve has an abrupt jump due to a shift to an alternative minimum of the GCV function. This finding is consistent with the simulation results of Section 9.3, and the fact that it occurred only for the larger numbers of components is unsurprising given that the degree of the polynomials of Section 8 increases linearly with the number of components.

10 Discussion

Smoothing parameter selection is often understood as steering a middle course between a model fit that is too close to the data, and one that is too smooth. This simplified picture suggests that the criterion optimized by selection methods such as GCV and REML should have a unique optimum. Sometimes, however, such functions can have multiple optima, leading to two possible problems. On the one hand, an algorithm seeking a unique optimum may in fact be finding a local, but not global, optimum. This phenomenon seems to be more common for GCV than for REML, but we have occasionally observed it when trying to maximize the latter criterion via mixed model software. On the other hand, the simulations in Section 9.3 suggest that the global optimum may not always represent the best choice: in particular, optima at low values of λ may perhaps reflect random fluctuations unrelated to the

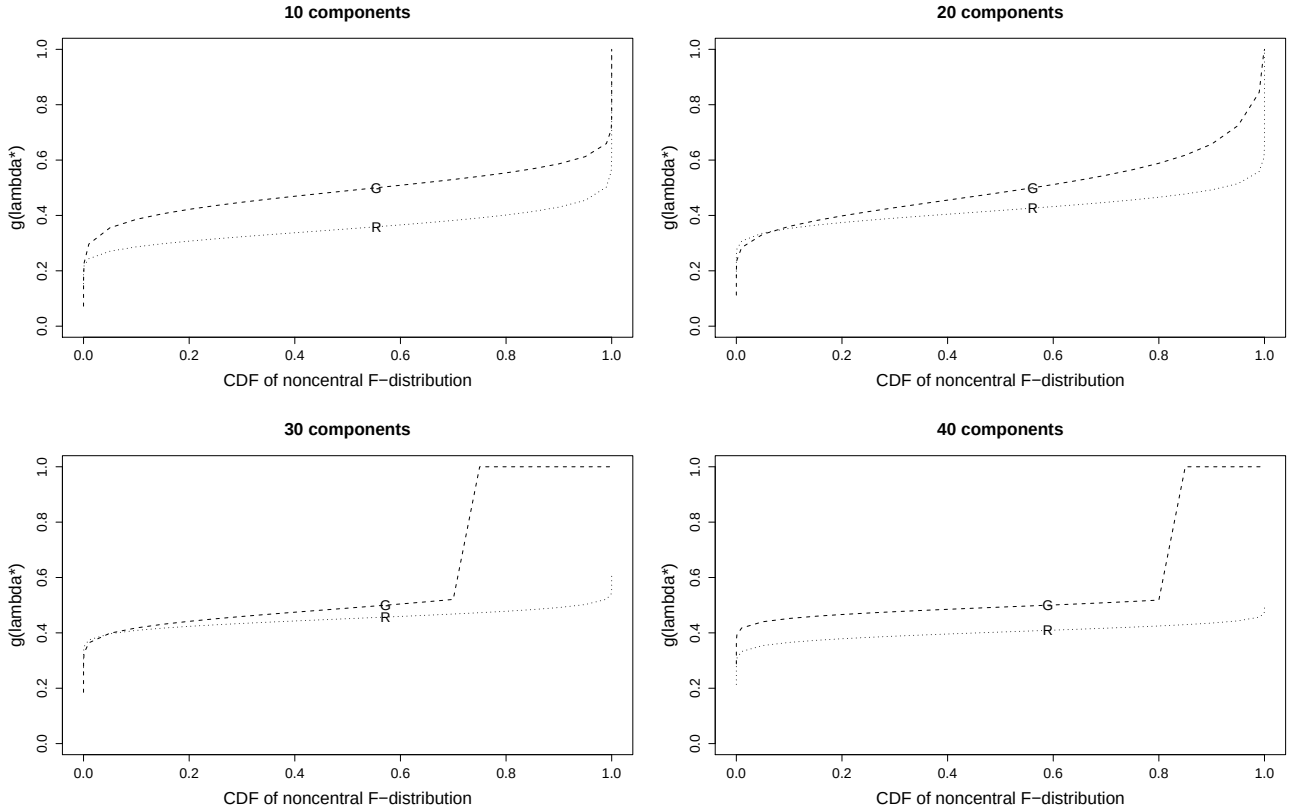

Figure 4: Sensitivity of λ to perturbations of the octane number outcome. The plotted function g (see text) illustrates how the choice of λ in FPCR models using perturbed outcomes (dotted curve: REML; dashed curve: GCV) compares with the value chosen by GCV with the actual octane numbers. 'R' and 'G' mark the results with the true outcome.

fidelity-smoothness tradeoff.

Practically speaking, therefore, we consider it advisable to plot the criteria over a wide range of plausible values of λ whenever possible, rather than automatically accepting any value at which criterion attains an optimum. The bounds on λ imposed by Theorem 1(i) may perhaps be exploited to facilitate this process, as may the relationship between stationary points of the GCV and REML criteria and the polynomials of Section 8. It may be a sound policy to compare results obtained by the two methods, and thereby use each as a check on the other. In some situations one could perhaps guard against optima at unreasonably small λ by imposing *a priori* a highest plausible number of degrees of freedom for the model fit.

The positive eigenvalues $\gamma_1, \ldots, \gamma_q$ of $\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T$ have played a key role in our arguments. Theorem 1(ii) suggests that variation in these eigenvalues contributes to instability in the choice of λ. In this sense an ideal situation occurs when $\gamma_1 = \ldots = \gamma_q$: then GCV and REML are equivalent, and both criteria are well-behaved provided $F_{(\mathbf{Z})} > 1$, as should usually be the case. These considerations lead us to speculate that, to the extent that one has a choice in the matter, it may be advantageous to use a design matrix $\boldsymbol{Z}$ in model (1) such that the positive eigenvalues of $\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T$ are as close together as possible. We plan to investigate this hypothesis in future work.

In many applications the function being estimated may have varying smoothness, so it would be preferable to have a locally adaptive smoothing parameter rather than the constant smoothing parameter we have assumed. Within the spline literature there have been several proposals for nonparametric regression with adaptive roughness penalties (e.g., Krivobokova, Crainiceanu and Kauermann, 2007, and references therein), and Cardot (2002) presents a method of this type for functional regression. Adaptive smoothing parameter selection strategies are generally "built upon" criteria such as GCV and REML in that they seek to optimize such criteria locally. Conse-

quently, in our experience, the pathologies that sometimes afflict constant-parameter smoothing can occur with adaptive smoothing as well. We therefore believe that a better understanding of how choice of a constant smoothing parameter works, and may fail to work, could help pave the way toward improved adaptive smoothing.

Appendix A: Derivative of the restricted likelihood with respect to λ

For given $\boldsymbol{\beta}$ and λ, the value of σ^2 maximizing the restricted log likelihood (7) is

$$\hat{\sigma}^2_{\beta,\lambda} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}_\lambda^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/(n-p); \tag{22}$$

substituting in this value and ignoring an additive constant leads to the profile restricted log likelihood

$$\begin{aligned} l_R(\boldsymbol{\beta}, \lambda | \boldsymbol{y}) &= -\frac{1}{2}\left[\log|\boldsymbol{V}_\lambda| + \log|\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|\right. \\ &\quad \left. +(n-p)\log\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{V}_\lambda^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\}\right]. \end{aligned} \tag{23}$$

For given λ, the value of $\boldsymbol{\beta}$ maximizing this last expression is the generalized least squares fit $\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{y}$.

Using the readily verified equality $\boldsymbol{V}_\lambda^{-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \lambda\boldsymbol{W})^{-1}\boldsymbol{Z}^T$, the following key facts about $\boldsymbol{P}_\lambda$ can be shown to hold under Assumptions 1–3:

$$\boldsymbol{P}_\lambda = \boldsymbol{I} - \boldsymbol{H}_\lambda, \tag{24}$$

where $\boldsymbol{H}_\lambda$ is the hat matrix defined by $\hat{\boldsymbol{y}} = \boldsymbol{H}_\lambda\boldsymbol{y}$ and given by

$$\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z} + \lambda\boldsymbol{W})^{-1}\boldsymbol{Z}^T; \tag{25}$$

$$\boldsymbol{V}_\lambda^{-1}\boldsymbol{X} = \boldsymbol{X}; \tag{26}$$

$$\boldsymbol{P}_\lambda^k = \boldsymbol{V}_\lambda^{-k} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \text{ for } k = 1, 2, \ldots. \tag{27}$$

Under Assumptions 1–3, repeated application of (26) gives $\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y}$, and hence

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)^T\boldsymbol{V}_\lambda^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda) = \boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y}. \tag{28}$$

Substituting (28) into (23) yields the profile restricted log likelihood for λ alone:

$$l_R(\lambda|\boldsymbol{y}) = -\frac{1}{2}\left[\log|\boldsymbol{V}_\lambda| + \log|\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}| + (n-p)\log(\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y})\right]. \tag{29}$$

Setting the derivative of (29) with respect of λ to zero will yield an equation for the REML estimate of λ. By (26) again, $\log|\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}| = \log|\boldsymbol{X}^T\boldsymbol{X}|$, which does not depend on λ, so the differentiation reduces to finding the derivatives of $\log|\boldsymbol{V}_\lambda|$ and $\log(\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y})$. To that end we shall need the (component-wise) derivatives of $\boldsymbol{V}_\lambda$ and $\boldsymbol{P}_\lambda$ with respect to λ; these can be shown to be:

$$\frac{d\boldsymbol{V}_\lambda}{d\lambda} = \lambda^{-1}(\boldsymbol{I} - \boldsymbol{V}_\lambda); \tag{30}$$

$$\frac{d\boldsymbol{P}_\lambda}{d\lambda} = \lambda^{-1}(\boldsymbol{P}_\lambda - \boldsymbol{P}_\lambda^2). \tag{31}$$

A formula in Lindstrom and Bates (1988, p. 1016), together with (30), leads to

$$\frac{d}{d\lambda}\log|\boldsymbol{V}_\lambda| = \lambda^{-1}\mathrm{tr}(\boldsymbol{V}_\lambda^{-1} - \boldsymbol{I}).$$

By (27), $\mathrm{tr}(\boldsymbol{V}_\lambda^{-1}) = \mathrm{tr}\boldsymbol{P}_\lambda + \mathrm{tr}[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] = \mathrm{tr}\boldsymbol{P}_\lambda + p$, so we conclude that

$$\frac{d}{d\lambda}\log|\boldsymbol{V}_\lambda| = \lambda^{-1}[\mathrm{tr}\boldsymbol{P}_\lambda - (n-p)]. \tag{32}$$

By Assumption 4, $\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y} > 0$. Thus, using (31), we obtain

$$\frac{d}{d\lambda}\log(\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y}) = \lambda^{-1}\left[1 - \frac{\boldsymbol{y}^T\boldsymbol{P}_\lambda^2\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{P}_\lambda\boldsymbol{y}}\right]. \tag{33}$$

By (29), (32) and (33), we obtain (8).

Appendix B: Eigendecomposition of $\boldsymbol{P}_\lambda$

In this appendix we identify a set of n orthonormal eigenvectors of $\boldsymbol{P}_\lambda$, and derive the associated eigenvalues as given in Section 6.

1. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_q$ are orthonormal eigenvectors corresponding to the positive eigenvalues $\gamma_1 \geq \ldots \geq \gamma_q$ of $\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T$; this is a maximal set of such eigenvectors since $\text{rank}(\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T) = q$. To see that these are also eigenvectors of $\boldsymbol{P}_\lambda$, observe that for $i \in 1, \ldots, q$, (6) leads to $\boldsymbol{V}_\lambda \boldsymbol{u}_i = (1 + \lambda^{-1}\gamma_i)\boldsymbol{u}_i$, so that

$$\boldsymbol{V}_\lambda^{-1}\boldsymbol{u}_i = (1 + \lambda^{-1}\gamma_i)^{-1}\boldsymbol{u}_i = \frac{\lambda}{\lambda + \gamma_i}\boldsymbol{u}_i;$$

and by Assumption 3, $\boldsymbol{X}^T\boldsymbol{u}_i = \gamma_i^{-1}\boldsymbol{X}^T\boldsymbol{Z}\boldsymbol{W}^{-1}\boldsymbol{Z}^T\boldsymbol{u}_i = \boldsymbol{0}$. Therefore

$$\begin{aligned} \boldsymbol{P}_\lambda \boldsymbol{u}_i &= [\boldsymbol{V}_\lambda^{-1} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{u}_i \\ &= \frac{\lambda}{\lambda + \gamma_i}\boldsymbol{u}_i, \end{aligned}$$

and $d_{\lambda i} = \frac{\lambda}{\lambda + \gamma_i}$ for $i = 1, \ldots, q$.

2. $\boldsymbol{u}_{q+1}, \ldots, \boldsymbol{u}_{n-p}$ are orthonormal vectors which, together with the p columns of $\boldsymbol{X}$, form a basis for null$(\boldsymbol{Z}^T)$. In view of (24), (25) and Assumption 3, we have $\boldsymbol{P}_\lambda \boldsymbol{u}_i = \boldsymbol{u}_i$ for $i = q + 1, \ldots, n - p$, and thus $d_{\lambda,q+1} = \ldots d_{\lambda,n-p} = 1$.

3. $\boldsymbol{u}_{n-p+1}, \ldots, \boldsymbol{u}_n$ form an orthonormal basis for the column space of $\boldsymbol{X}$. Using (24), (25) and Assumption 3 again, we obtain $\boldsymbol{P}_\lambda \boldsymbol{u}_i = \boldsymbol{0}$ for $i = n-p+1, \ldots, n$, and so $d_{\lambda,n-p+1} = \ldots = d_{\lambda n} = 0$.

Appendix C: Proofs of Theorems 1–3

Proof of Theorem 1

We shall need the following elementary result, whose proof is omitted.

Lemma 1 *Consider weighted averages a_1, a_2 given by $a_k = \sum_{i=1}^{g+1} w_{kj} s_j \,/ \sum_{i=1}^{g+1} w_{kj}$, $k = 1, 2$, where $w_{1i} > w_{2i} > 0$ for $i = 1, \ldots, g$ and $w_{1,g+1} = w_{2,g+1} > 0$.*

(i) If $s_i \leq s_{g+1}$ for $i = 1, \ldots, g$, with strict inequality for some i, then $a_1 < a_2$.

(ii) If $s_i \geq s_{g+1}$ for $i = 1, \ldots, g$, with strict inequality for some i, then $a_1 > a_2$.

(iii) If $s_1 = \ldots = s_g = s_{g+1}$ then $a_1 = a_2$.

(i) In view of the discussion in Section 6, it suffices to show that if $\sum_{i=q+1}^{n-p} y_{(i)}^2 = 0$ then $h_1(\lambda) > h_2(\lambda) > h_3(\lambda)$ for all $\lambda > 0$. The proof relies on expressing $h_m(\lambda)$ $(m = 1, 2, 3)$ as a weighted average as in the text at (17). To see that $h_1(\lambda) > h_2(\lambda)$ for all $\lambda > 0$, apply Lemma 1(ii) with $g = q$, $s_i = y_{(i)}^2 d_{\lambda j}$ and $w_{ki} = d_{\lambda 1}^{k-1}$ for $i = 1, \ldots, q$ and $k = 1, 2$, $s_{q+1} = 0$, and $w_{q+1} = n - p - q$. (Note that in this case Assumption 4 ensures that $s_i > s_{q+1}$ for some $i \in \{1, \ldots, q\}$, as is required in order for Lemma 1(ii) to apply.) The proof that $h_2(\lambda) > h_3(\lambda)$ for all $\lambda > 0$ is identical except that we take $w_{ki} = d_{\lambda 1}^{k}$.

(ii) Suppose $\sum_{i=q+1}^{n-p} y_{(i)}^2 > 0$.

(a) If $F_{(i)} \leq 1$ for $i = 1, \ldots, q$ then for all $\lambda > 0$, the first q quantities in (17) are less than the last and thus (15) holds, by Lemma 1(i). Hence both GCV and REML choose $\lambda = \infty$.

(b) Suppose $0 < \lambda < \min_{\{1 \leq i \leq q,\ F_{(i)} > 1\}} \frac{\gamma_i}{F_{(i)} - 1}$. Separate consideration of the $F_{(i)} > 1$ and $F_{(i)} \leq 1$ cases confirms that the first q of the $q + 1$ quantities listed in (17) are smaller than the $(q+1)$th. By Lemma 1(i) again, we obtain that (15) holds on $(0, \min_{\{1 \leq i \leq q,\ F_{(i)} > 1\}} \frac{\gamma_i}{F_{(i)} - 1})$, and the lower bound follows.

(c) If $F_{(i)} > 1$ for $i = 1, \ldots, q$ then for $\lambda > \max_{1 \leq i \leq q} \frac{\gamma_i}{F_{(i)} - 1}$, the first q elements listed in (17) are larger than the $(q + 1)$th. By Lemma 1(ii), (16) holds for $\lambda > \max_{1 \leq i \leq q} \frac{\gamma_i}{F_{(i)} - 1}$, and the upper bound follows. □

Proof of Theorem 2

For $m = 0, 1, \ldots$ and $i = 1, \ldots, q$ we have $d_{\lambda i}^m = 1 - m e_{\lambda i} + R_{\lambda i m}$, where $R_{\lambda i m}$ is a linear combination of second and higher powers of $e_{\lambda i} = \frac{\gamma_i}{\lambda + \gamma_i}$. Thus $d_{\lambda i}^m = 1 - m e_{\lambda i} + o(\lambda^{-1})$ and, from (34),

$$\begin{aligned} t_{\lambda,k-1} Q_{\lambda,k+1} - t_{\lambda k} Q_{\lambda k} &= \left[n - p - (k-1) \sum_{i=1}^{q} e_{\lambda i} + o(\lambda^{-1}) \right] \\ &\quad \times \left[\sum_{i=1}^{q} y_{(i)}^2 \{1 - (k+1) e_{\lambda i} + o(\lambda^{-1})\} + \sum_{i=q+1}^{n-p} y_{(i)}^2 \right] \\ &\quad - \left[n - p - k \sum_{i=1}^{q} e_{\lambda i} + o(\lambda^{-1}) \right] \\ &\quad \times \left[\sum_{i=1}^{q} y_{(i)}^2 \{1 - k e_{\lambda i} + o(\lambda^{-1})\} + \sum_{i=q+1}^{n-p} y_{(i)}^2 \right] \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{q} e_{\lambda i} \left[\sum_{i=1}^{q} y_{(i)}^2 \{1 - (k+1) e_{\lambda i}\} + \sum_{i=q+1}^{n-p} y_{(i)}^2 \right] \\ &\quad - \left(n - p - k \sum_{i=1}^{q} e_{\lambda i} \right) \sum_{i=1}^{q} y_{(i)}^2 e_{\lambda i} + o(\lambda^{-1}) \end{aligned}$$

$$= \sum_{i=1}^{q} e_{\lambda i} \sum_{i=1}^{n-p} y_{(i)}^2 - (n-p) \sum_{i=1}^{q} y_{(i)}^2 e_{\lambda i} + o(\lambda^{-1}).$$

Ignoring the $o(\lambda^{-1})$ term, the last expression is negative if $\frac{\sum_{i=1}^{n-p} y_{(i)}^2}{n-p} < \frac{\sum_{i=1}^{q} y_{(i)}^2 e_{\lambda i}}{\sum_{i=1}^{q} e_{\lambda i}}$, and the right side of the preceding converges to $\frac{\sum_{i=1}^{q} y_{(i)}^2 \gamma_i}{\sum_{i=1}^{q} \gamma_i}$ as $\lambda \to \infty$. The theorem follows by (13) and (14). □

Proof of Theorem 3

For $m = 1, 2, 3$, $h_m(\lambda)(n-p-q) / \sum_{i=q+1}^{n-p} y_{(i)}^2$ is a weighted average of $\frac{\lambda F_{(\mathbf{Z})}}{\lambda + \gamma}$ and 1 with weights $q(\frac{\lambda}{\lambda+\gamma})^{m-1}$, $n-p-q$.

(i) If $F_{(\mathbf{Z})} > 1$ then $\lambda = \frac{\gamma}{F_{(\mathbf{Z})} - 1}$ leads to $\frac{\lambda F_{(\mathbf{Z})}}{\lambda + \gamma} = 1$, so Lemma 1(iii) applies and

$h_1(\lambda) = h_2(\lambda) = h_3(\lambda)$. By Lemma 1(i)-(ii), (15) and (16) hold for smaller and larger λ, respectively. Thus the given λ is the unique optimum for both criteria.

(ii) If $F_{(\mathbf{Z})} \leq 1$ then, for all $\lambda > 0$, $\frac{\lambda F_{(\mathbf{Z})}}{\lambda+\gamma} < 1$ and hence (15) holds by Lemma 1(i). It follows that $\lambda = \infty$ is the unique optimum for both GCV and REML. □

Appendix D: The polynomials of Section 8

By (11) and (12),

$$\begin{aligned} t_{\lambda,k-1}Q_{\lambda,k+1} - t_{\lambda k}Q_{\lambda k} &= \left(n-p-q+\sum_{i=1}^{q} d_{\lambda i}^{k-1}\right)\left(\sum_{i=1}^{q} y_{(i)}^2 d_{\lambda i}^{k+1} + \sum_{i=q+1}^{n-p} y_{(i)}^2\right) \\ &\quad - \left(n-p-q+\sum_{i=1}^{q} d_{\lambda i}^{k}\right)\left(\sum_{i=1}^{q} y_{(i)}^2 d_{\lambda i}^{k} + \sum_{i=q+1}^{n-p} y_{(i)}^2\right) \quad (34) \end{aligned}$$

Let $g_i(\theta) = d_{\lambda i}^{-1} = 1 + \gamma_i\theta$ where $\theta = 1/\lambda$. Some more algebra shows that, for $\theta \neq 0$,

$$\begin{aligned} (t_{\lambda,k-1}Q_{\lambda,k+1} - t_{\lambda k}Q_{\lambda k})\prod_{i=1}^{q} g_i(\theta)^{2k}/\theta &= \sum_{i=q+1}^{n-p} y_{(i)}^2 \sum_{i=1}^{q}\left[\gamma_i g_i(\theta)^k \prod_{j\neq i} g_j(\theta)^{2k}\right] \quad (35) \\ &\quad -(n-p-q)\sum_{i=1}^{q}\left[\gamma_i y_{(i)}^2 g_i(\theta)^{k-1}\prod_{j\neq i} g_j(\theta)^{2k}\right] \\ &\quad +\sum_{i=1}^{q}\prod_{j\neq i} g_j(\theta)^{k-1}\sum_{i=1}^{q}\left[y_{(i)}^2\prod_{j\neq i} g_j(\theta)^k \frac{\prod_{j\neq i} g_j(\theta)-1}{\theta}\right] \\ &\quad -\sum_{i=1}^{q}\left[\prod_{j\neq i} g_j(\theta)^{k-1}\frac{\prod_{j\neq i} g_j(\theta)-1}{\theta}\right]\sum_{i=1}^{q}\left[y_{(i)}^2\prod_{j\neq i} g_j(\theta)^k\right], \end{aligned}$$

where $\prod_{j\neq i}$ is shorthand for the product over all $j \in \{1,\ldots,q\}$ except i. This is a polynomial in θ of degree $k(2q-1)$, the positive zeros of which are the reciprocals of the positive stationary points of $l_R(\lambda)$ or $GCV(\lambda)$. Thus, taking $k = 1$, we find that the number of positive stationary points of $l_R(\lambda)$ is the number of positive roots of a particular $(2q-1)$th-degree polynomial; taking $k = 2$, the number of stationary points of $GCV(\lambda)$ is the number of positive roots of a polynomial of degree $4q-2$. Some further algebra reveals that when $k > 1$, polynomial (35) has negative roots $-\gamma_1^{-1},\ldots,-\gamma_q^{-1}$; thus we can replace $4q-2$ with $3q-2$ in the previous statement.

References

1. Cantoni, E., and Hastie, T. (2002) Degrees-of-freedom tests for smoothing splines. *Biometrika*, 89, 251–263.

2. Cardot, H. (2002) Spatially adaptive splines for statistical linear inverse problems. *Journal of Multivariate Analysis*, 81, 100–119.

3. Cardot, H., Ferraty, F., and Sarda, P. (2003) Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571–591.

4. Cox, D., Koh, E., Wahba, G., and Yandell, B. S. (1988) Testing the (parametric) null hypothesis in (semiparametric) partial and generalized spline models. *Ann. Statist.*, 16, 113–119.

5. Crainiceanu, C., and Ruppert, D. (2004) Likelihood ratio tests in linear mixed models with one variance component. *J. R. Statist. Soc.* B, 66, 165–185.

6. Craven, P., and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31, 377–403.

7. Eilers, P. H. C. (1999) Discussion of "The analysis of designed experiments and longitudinal data by using smoothing splines" by A. P. Verbyla, B. R. Cullis, M. G. Kenward and S. J. Welham. *J. R. Statist. Soc.* C, 307–308.

8. Eilers, P. H. C. and Marx, B. D., (2002) Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, 11, 758–783.

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

10. Hastie, T. J., and Tibshirani, R. J. (1990) *Generalized Additive Models*. Boca Raton: Chapman & Hall/CRC.

11. Kalivas, J. H. (1997) Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37, 255–259.

12. Kauermann, G. (2005) A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127, 53–69.

13. Kou, S. C. (2004). From finite sample to asymptotics: a geometric bridge for selection criteria in spline regression. *Ann. Statist.*, 32, 2444–2468.

14. Krivobokova, T., Crainiceanu, C. M., Kauermann, G. (2008). Fast adaptive penalized splines, *Journal of Computational and Graphical Statistics*, 17, 1–20.

15. Lee, Y., Nelder, J. A., and Pawitan, Y. (2006) *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihoood*. Boca Raton: Chapman & Hall/CRC.

16. Lindstrom, M. J., and Bates, D. M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Statist. Ass.*, 83, 1014–1022.

17. Marx, B. D., and Eilers, P. H. C. (1999) Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, 41, 1–13.

18. Pawitan, Y. (2001) *In All Likelihood*. Oxford: Oxford University Press.

19. Reiss, P. T. (2006) Regression with signals and images as predictors. Ph.D. dissertation, Department of Biostatistics, Columbia University.

20. Reiss, P. T., and Ogden, R. T. (2007) Functional principal component regression and functional partial least squares. *J. Am. Statist. Ass.*, 102, 984–996.

21. Ruppert, D., Wand, M. P., and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge and New York: Cambridge University Press.

22. Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.