Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Statistical Planning and Inference 127 (2005) 53–69

journal of statistical planning and inference

www.elsevier.com/locate/jspi

A note on smoothing parameter selection for penalized spline smoothing

Göran Kauermann

Department of Economics, University Bielefeld, Postfach 300131, Bielefeld 33501, Germany

Received 15 November 2002; accepted 6 September 2003

Abstract

In nonparametric regression the smoothing parameter can be selected by minimizing a Mean Squared Error (MSE) based criterion. For spline smoothing one can also rewrite the smooth estimation as a Linear Mixed Model where the smoothing parameter appears as the a priori variance of spline basis coefficients. This allows to employ Maximum Likelihood (ML) theory to estimate the smoothing parameter as variance component. In this paper the relation between the two approaches is illuminated for penalized spline smoothing (*P*-spline) as suggested in Eilers and Marx Statist. Sci. 11(2) (1996) 89. Theoretical and empirical arguments are given showing that the ML approach is biased towards undersmoothing, i.e. it chooses a too complex model compared to the MSE. The result is in line with classical spline smoothing, even though the asymptotic arguments are different. This is because in *P*-spline smoothing a finite dimensional basis is employed while in classical spline smoothing the basis grows with the sample size.

MSC: 62G08; 62G20; 62J07

Keywords: Linear Mixed Model; *P*-spline smoothing; REML estimate; Smoothing; Smoothing parameter selection

1. Introduction

Penalized spline estimation (*P*-spline) for smoothing traces back to Parker and Rice (1985) and O'Sullivan (1986), but it was Eilers and Marx (1996) who made the method popular by illuminating the numerical practicability and flexibility of the approach. The major idea behind *P*-spline estimation is thereby simple. For smooth estimation a large

E-mail address: gkauermann@wiwi.uni-bielefeld.de (G. Kauermann).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.09.023

but finite dimensional basis is employed. Instead of simple parametric fitting, however, which would lead to variable and wiggling estimates, a penalized version is pursued to provide a smooth fit. Practical experience has shown that the concrete specification of the basis and its dimension has little influence on the fit (see e.g. French et al., 2001 or Ruppert, 2002). More relevant for the smoothness of the fit is the amount of penalization applied. Some first theoretical considerations on how to choose the right amount of penalization are found in Wand (1999) or Aerts et al. (2002).

The idea of P-spline smoothing is strongly related to Linear Mixed Models. This becomes obvious if the basis coefficients are considered as random effects and the penalization appears as a priori distribution imposed on the basis coefficients. In this scenario spline smoothing is equivalent to (maximum) posterior Bayes estimation in the resulting Linear Mixed Model and the smoothing parameter plays the role of the a priori variance of the basis coefficients. This in turn can be estimated using Maximum Likelihood (ML) theory as suggested in Wecker and Ansley (1983) and further discussed, e.g. in Wahba (1985), Li (1985), Stein (1990) or Speckman and Sun (2001). Recently, Efron (2001) and Kou and Efron (2002) illuminate the connection from a geometrical point of view. References discussing the relation between spline smoothing and Mixed Models in general include also Green and Silverman (1994), Brumback and Rice (1998) or Verbyla et al. (1999).

For P-spline smoothing there appears a major difference compared to classical spline smoothing treated in the above-cited papers. In classical spline smoothing for each observation a separate basis function is included. This means that the resulting basis matrix is $n \times n$ dimensional, with n as sample size. In contrast, for P-spline smoothing a prespecified high but finite dimensional basis is used. This allows to exploit the link to Linear Mixed Models not only from a theoretical angle but also practically. In particular Linear Mixed Models software (see e.g. Pinheiro and Bates, 2000) can be used for smoothing (see Wand, 2003). In case of a non-normal response this generalizes to Generalized Linear Mixed Models with penalized quasi-likelihood estimation (see also Breslow and Clayton, 1993). Again, in the Linear Mixed Model formulation the smoothing parameter steering the amount of smoothing is the ratio of the a priori variance of the basis coefficients and the residual variance. This in turn suggests to take the ML or the Residual Maximum Likelihood (REML) estimator (Harville, 1977) as smoothing parameter selection.

This note intends to illuminate the REML choice in more depth. We show that asymptotically REML-based smoothing parameter selection is biased towards undersmoothing. This resembles results found for spline smoothing (see e.g. Efron, 2001). Our asymptotic arguments are however different to those used in classical spline smoothing. This is since for P-spline smoothing a finite dimensional basis is used while for classical spline smoothing the basis grows with the sample size. This also implies that asymptotically P-spline smoothing leads to standard parametric fitting and penalization is losing its effect for growing sample size.

The paper is organized as follows. In Section 2, we introduce different Smoothing parameter selection routines. Asymptotic investigation is provided in Section 3 while Section 4 explores the finite sample performance. A discussion concludes the paper.

2. Smoothing and mixed models

2.1. Mean-squared error

Let us consider the simple smoothing model

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

with x_i as metrical covariate, $f(\cdot)$ as unknown smooth function and ε_i as independent normally distributed residuals, i.e. $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. For simplicity of presentation we will assume that σ_ε^2 is known. P-spline estimation is now pursued by replacing $f(\cdot)$ by the parametric form

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{b} + \varepsilon_i, \tag{2}$$

where $\boldsymbol{x}_i$ is a low-dimensional parametric basis build from x_i, e.g. the linear basis $\boldsymbol{x}_i = (1, x_i)^{\mathrm{T}}$, and $\boldsymbol{z}_i$ is a high-dimensional basis linearly independent of x_i. A convenient choice is to use truncated polynomials, i.e. $\boldsymbol{z}_i = \{(x_i - \tau_1)_+, \ldots, (x_i - \tau_K)_+\}$ with $(\cdot)_+$ as positive part, that is $(x)_+ = x$ for $x \geqslant 0$ and $(x)_+ = 0$ for $x < 0$. The knots τ_k are thereby fixed values covering the range of x_i, $i = 1, \ldots, n$.

The general idea is to choose basis $\boldsymbol{z}_i$ in a "lush" and "generous" manner such that the difference $\delta(x_i) = f(x_i) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{b}$ is of ignorable size. In particular we assume that dimension K of basis $\boldsymbol{z}$ is large but finite and fixed independently of sample size n. Direct estimation of coefficients $\boldsymbol{\beta}$ and $\boldsymbol{b}$ by maximizing the likelihood resulting from (2) would lead to highly variable and wiggled estimates for $f(x)$. To achieve smoothness a penalty is introduced leading to the penalized likelihood

$$l(\beta; b; \lambda) = -\tfrac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}) - \tfrac{1}{2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{b}/\lambda \tag{3}$$

with λ as smoothing parameter and $\boldsymbol{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$ and analogously $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^{\mathrm{T}}$. Matrix $\boldsymbol{D}_K$ is a $K \times K$ dimensional penalty matrix which for reasons to become clear later is assumed to be symmetrical and invertible. Differentiating (3) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ leads to the estimating equations:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{b}}), \tag{4}$$

$$\hat{\boldsymbol{b}} = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{D}_K/\lambda)^{-1}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{5}$$

Parameter λ in (3) plays the role of a smoothing parameter. Letting λ tend to infinity leads to standard maximum likelihood estimates while $\lambda \to 0$ implies $\hat{\boldsymbol{b}} \to 0$ so that $\hat{f}(x) = \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ results as parametric fit. A reasonable choice for λ is obtained by minimizing the Mean-Squared Error (MSE). We therefore assume that covariates x_i have compact support so that Fisher information matrices are of order $\mathrm{O}(n)$. This means for instance that matrix $\boldsymbol{F}_{Z.X} := n(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} - \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z})^{-1}$ has order $\mathrm{O}(1)$. Similar to the results provided in Wand (1999) we get the optimal MSE smoothing parameter

$$\lambda_{\mathrm{MSE}} = \frac{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K\boldsymbol{b} + 3\sigma_\varepsilon^2\,\mathrm{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)/n}{\sigma_\varepsilon^2\,\mathrm{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)} + \mathrm{O}(n^{-2}) \tag{6}$$

with tr(·) denoting the trace of a matrix. A short sketch of this statement is provided in the appendix. The optimal MSE smoothing parameter depends on both, the unknown but fixed coefficient $\boldsymbol{b}$ and matrix $\boldsymbol{F}_{ZX}$. It is also worth noting that for $\sigma_\varepsilon > 0$, λ_{MSE} has order O(1).

2.1.1. Cp estimate

An asymptotically unbiased estimate for the optimal MSE smoothing parameter is obtained, for instance, from the Cp criterion (see Mallows, 1973)

$$C_{\mathrm{Cp}}(\lambda) = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{Z}\hat{\boldsymbol{b}}) + 2\sigma_\varepsilon^2 \operatorname{tr}(\boldsymbol{S}_\lambda), \tag{7}$$

where $\boldsymbol{S}_\lambda$ is the smoothing matrix defined via $\boldsymbol{S}_\lambda \boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}}$, see the appendix for details. Minimizing (7) provides the smoothing parameter estimate $\hat{\lambda}_{\mathrm{Cp}}$. Straightforward calculation reveals the asymptotic form

$$\hat{\lambda}_{\mathrm{Cp}} = \frac{\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{ZX}\boldsymbol{D}_K\hat{\boldsymbol{b}}}{\sigma_\varepsilon^2 \operatorname{tr}(\boldsymbol{F}_{ZX}\boldsymbol{D}_K)}\{1 + \mathrm{O}_{\mathrm{p}}(n^{-1})\}, \tag{8}$$

so that $\hat{\lambda}_{\mathrm{Cp}}$ results as plug-in estimate of (6). For practical purposes form (8) is of little use unless the sample size is very large. Therefore a grid search to minimize (7) should be preferred.

2.2. REML estimate

The penalized likelihood (3) resembles the likelihood in the Linear Mixed Model

$$\boldsymbol{b} \sim \mathrm{N}(0, \sigma_b^2 \boldsymbol{D}_K^{-1}), \qquad \boldsymbol{Y}|\boldsymbol{b} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma_\varepsilon^2 \boldsymbol{I}_n). \tag{9}$$

Considering $\boldsymbol{b}$ as random effect we can marginalize (9) and get

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{V}_\lambda) \tag{10}$$

with $\boldsymbol{V}_\lambda = \boldsymbol{I}_n + \lambda \boldsymbol{Z}\boldsymbol{D}_K^{-1}\boldsymbol{Z}^{\mathrm{T}}$ and $\lambda = \sigma_b^2/\sigma_\varepsilon^2$. Model (10) is well established (see e.g. Searle et al., 1992) and the best linear unbiased predictor for $\boldsymbol{b}$ is given by (5). It is classical theory that (4) gives to the maximum likelihood estimate for $\boldsymbol{\beta}$ and simple algebra leads to the form

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}\boldsymbol{Y}). \tag{11}$$

In model (10) the smoothing parameter λ relates to the a priori variance of $\boldsymbol{b}$. This can be estimated by maximizing the REML likelihood (see Harville, 1977)

$$l_{\mathrm{REML}}(\beta; \lambda) = -\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{\sigma_\varepsilon^2} - \log|\boldsymbol{V}_\lambda| - \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}|. \tag{12}$$

Differentiating (12) with respect to λ and inserting estimates for $\boldsymbol{\beta}$ provides the REML estimate (see the appendix for details)

$$\hat{\lambda}_{\mathrm{REML}} = \frac{\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\hat{\boldsymbol{b}}/\sigma_\varepsilon^2 + \operatorname{tr}(\boldsymbol{F}_{ZX}\boldsymbol{D}_K/n)}{K} + \mathrm{O}(n^{-2}). \tag{13}$$

In general, REML estimation takes the estimation of $\boldsymbol{\beta}$ into account which is mirrored in the latter component in (12) and likewise in the second component in the numerator in (13). If instead simple maximum likelihood estimation of λ is pursued the second component of the numerator of (13) is omitted.

3. Asymptotic comparison of smoothing parameter selectors

3.1. Mixed model

We will now compare the two smoothing parameter selectors $\hat{\lambda}_{\mathrm{REML}}$ and λ_{MSE}. The conceptional difference between the two approaches is obvious. For the REML estimate we assume the Linear Mixed Model (9) to hold, in particular coefficient vector $\boldsymbol{b}$ is considered as random. In contrast for MSE smoothing parameter selection we take $\boldsymbol{b}$ as fixed but unknown. This means we assume a model which consists of the second part in (9) only, that is we condition on $\boldsymbol{b}$. The following theorem illuminates the behavior of the smoothing parameter estimates if the Linear Mixed Model (9) holds and $\boldsymbol{b}$ is random.

Theorem. *Assuming x_i to have compact support and considering model (9) as true model we get asymptotically*

$$P(\hat{\lambda}_{\mathrm{REML}} > \lambda_{\mathrm{MSE}}) = P\left(\sum_{k=1}^{K} v_k \mathcal{X}_k^2 > 0\right) + \mathrm{O}(n^{-1}), \tag{14}$$

where $\mathcal{X}_k^2$, $k=1,\ldots,K$ are independent Chi squared distributed variables with 1 degree of freedom and $v_k = 1/K - \rho_k / \sum_{l=1}^{K} \rho_l$, where ρ_k are the eigenvalues of $\boldsymbol{F}_{ZX}\boldsymbol{D}_K$.

The proof of the theorem is provided in the appendix. In principle $P(\sum_{k=1}^{K} v_k \mathcal{X}_k^2 > 0)$ can be calculated using the ideas of Davies (1980), even though nowadays simple numerical simulation techniques appear more natural. The result gives the asymptotic probability that the REML estimate undersmoothes. In standard scenarios this probability will be larger than 0.5. For instance for truncated polynomials taking $\boldsymbol{D}_K = \boldsymbol{I}_K$ the eigenvalues of $\boldsymbol{F}_{ZX}$ are skewly distributed so that $\sum_{k=1}^{K} (1/K - \rho_k / \sum_l \rho_l)^3 < 0$. This in turn implies $P(\sum_k v_k \mathcal{X}_k^2 > 0) > 0.5$. The simulation study below illuminates this point in more depth.

3.1.1. Simulation study

We run a small simulation study to visualize the above result. A more comprehensive study focusing on small sample properties is given in the next section. We draw $n=250$ and 750 data points from the model $y_i = (1, x_i)\boldsymbol{\beta} + \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{b} + \varepsilon_i$ with $\boldsymbol{\beta} = \boldsymbol{0}$ for simplicity, and x_i as equidistant points on $[0,1]$. Basis $\boldsymbol{z}$ is built from $K = 30$ truncated linear lines with $\boldsymbol{D}_K = \boldsymbol{I}_K$ (as has been suggested as penalty matrix for this basis by Ruppert and Carroll, 2000). Components $\boldsymbol{b}$ are drawn independently from a standard normal distribution while $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.2$. Smoothing parameter estimates are found

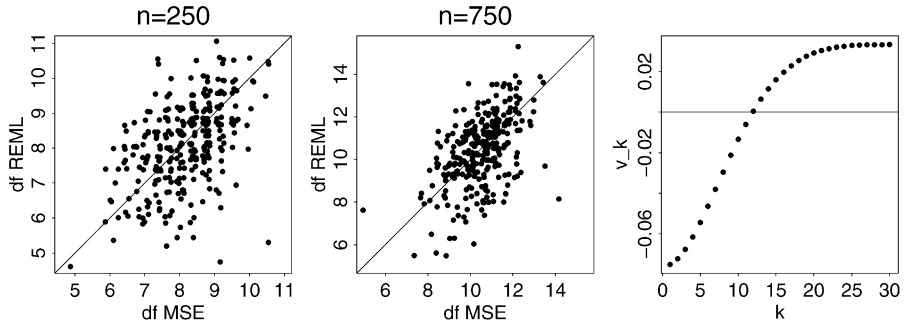

Fig. 1. Degrees of freedom $\mathrm{df}(\lambda_{\mathrm{MSE}})$ plotted against $\mathrm{df}(\hat{\lambda}_{\mathrm{REML}})$ for $n=250$ (upper plot) and $n=750$ (middle plot). Coefficients v_k (bottom plot), indistinguishable for $n=250$ and 750.

Table 1
Empirical and asymptotic probability for undersmoothing

	$P(\hat{\lambda}_{\mathrm{REML}} > \lambda_{\mathrm{MSE}})$	
	$n=250$	$n=750$
Empirical	50.4	53.0
Asymptotic	50.9	54.1

based on a 50 dimensional grid search. Note that λ_{MSE} is calculated conditional on $\boldsymbol{b}$, so that by simulating $\boldsymbol{b}$ it is random.

Fig. 1 shows the results based on 300 simulations. Instead of plotting the actual values of λ, which are hard to interpret, we plot the corresponding degrees of freedom, that is $\mathrm{df}(\lambda)=\mathrm{tr}(\boldsymbol{S}_\lambda)$. The left plot is for $n=250$, the middle plot for $n=750$ (points have been jittered for better visual impression). The scatterplots of $\mathrm{df}(\hat{\lambda}_{\mathrm{REML}})$ against $\mathrm{df}(\lambda_{\mathrm{MSE}})$ show a reasonable amount of correlation. We are however interested in the proportion of points lying above the diagonal. This is summarized in Table 1 (with choices $\lambda_{\mathrm{MSE}}=\hat{\lambda}_{\mathrm{REML}}$ due to the grid search divided uniformly on the two groups $\hat{\lambda}_{\mathrm{REML}} < \lambda_{\mathrm{MSE}}$ and $\hat{\lambda}_{\mathrm{REML}} > \lambda_{\mathrm{MSE}}$, respectively). As can be seen the REML estimate is undersmoothing as stated in the theorem, even though the effect is weak but increases with growing sample size. To complete the picture we also calculate the asymptotic distribution based on (14) by simulating $\sum_{k=1}^{K} v_k \chi^2_k$. Coefficients v_k are shown in the right plot of Fig. 1 and the corresponding simulated probabilities are included in Table 1. The simulations clearly support the theoretical findings.

3.2. Smoothing model

The result above is derived under the assumption of a Linear Mixed Model, that is coefficient vector $\boldsymbol{b}$ is assumed to be normally distributed. A more realistic scenario for

smoothing is however to consider $\boldsymbol{b}$ as fixed but unknown. This means we assume $f(x)$ to be approximated by $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$, where coefficients $\boldsymbol{\beta}$ and $\boldsymbol{b}$ are unknown but fixed. The optimal MSE smoothing parameter can then be estimated using the Cp criterion (7) and from (8) we get the asymptotic relationship

$$\hat{\lambda}_{\mathrm{REML}} - \hat{\lambda}_{\mathrm{Cp}} = \frac{1}{\sigma_\varepsilon^2}\,\hat{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{D}_K/K - \boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K/\mathrm{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K))\hat{\boldsymbol{b}}\{1 + \mathrm{O_p}(n^{-1})\}.$$

For simplicity we set $\boldsymbol{D}_K = \boldsymbol{I}_K$ subsequently and let as above ρ_k be the eigenvalues of $\boldsymbol{F}_{Z.X}$ corresponding to the eigenvectors $\boldsymbol{u}_k$, $k = 1,\ldots,K$. This yields $\boldsymbol{U} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_K)$ as resulting eigenbasis and for $\hat{\boldsymbol{c}} = \boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{b}}$ we get the form

$$\hat{\lambda}_{\mathrm{REML}} - \hat{\lambda}_{\mathrm{Cp}} = \frac{1}{\sigma_\varepsilon^2}\,\hat{\boldsymbol{c}}^{\mathrm{T}}\left(\boldsymbol{I}_K/K - \mathrm{diag}(\rho_k)\Big/\sum_l \rho_l\right)\hat{\boldsymbol{c}}\{1 + \mathrm{O_p}(n^{-1})\}. \tag{15}$$

Note that estimate $\hat{\boldsymbol{b}}$ is asymptotically $\mathrm{N}(\boldsymbol{b}, \sigma_\varepsilon^2\boldsymbol{F}_{Z.X}/n)$ distributed assuming $\lambda = \mathrm{O}(1)$ (see also (29) in the appendix). Accordingly we get asymptotically $\hat{\boldsymbol{c}} \sim \mathrm{N}(\boldsymbol{c}, \sigma_\varepsilon^2\mathrm{diag}(\rho_k)/n)$, with $\boldsymbol{c} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{b}$, so that the quadratic form (15) allows for the approximation

$$\hat{\lambda}_{\mathrm{REML}} - \hat{\lambda}_{\mathrm{Cp}} \sim \frac{1}{\sigma_\varepsilon^2 n}\sum_{k=1}^{K}\gamma_k\chi^2_{k,\delta} \tag{16}$$

with χ^2_{k,δ_k} as noncentral Chi-squared variables and $\gamma_k = \rho_k/K - \rho_k^2/\sum_l \rho_l$, $k = 1,\ldots,K$. The noncentrality parameters δ_k result from the fact that $\boldsymbol{b}$ is assumed to be fixed and not necessarily zero. Apparently, if $\boldsymbol{b} = \boldsymbol{0}$ the noncentrality vanishes and one obtains a behavior similar to (14) in the above theorem. Considering the noncentrality in more depth reveals the bias

$$E(\hat{\lambda}_{\mathrm{REML}} - \hat{\lambda}_{\mathrm{Cp}}) = \frac{1}{\sigma_\varepsilon^2}\left(\frac{\sum_k c_k^2}{K} - \frac{\sum_k c_k^2\rho_k}{\sum_k \rho_k}\right) + \mathrm{O}(n^{-1}), \tag{17}$$

where $\boldsymbol{c} = (c_1,\ldots,c_K)^{\mathrm{T}}$. In applications this bias will be typically positive meaning that $\hat{\lambda}_{\mathrm{REML}}$ is biased toward undersmoothing. To demonstrate this point we consider the transformed basis $\boldsymbol{Z}_X\boldsymbol{U}$, where $\boldsymbol{Z}_X = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}})\boldsymbol{Z}$. Note that $\boldsymbol{Z}_X\boldsymbol{b} = \boldsymbol{Z}_X\boldsymbol{U}\boldsymbol{c}$, so that $\boldsymbol{c}$ results as the coefficient vector for the transformed basis $\boldsymbol{Z}_X\boldsymbol{U}$. The columns of the transformed basis matrix relate to the eigenvalues ρ_k such that the larger eigenvalue ρ_k the more complex is the basis function given by the kth column of $\boldsymbol{Z}_X\boldsymbol{U}$, $k = 1,\ldots,K$. Using a 30 dimensional basis built from truncated linear lines we show in Fig. 2 for two different underlying functions (see plots in left column) the corresponding coefficient c_k (plots in middle column) in decreasing order of the eigenvalues. Bias (17) is mirrored in the right column where we show c_k^2/K plotted against $c_k^2\rho_k/\sum_k \rho_k$. All points lie below the diagonal which means that quantity (17) is positive. Consequently, the REML estimate is asymptotically biased and will undersmooth. Such behavior can be observed as long as the true underlying function can be well approximated by basis functions in $\boldsymbol{Z}_X\boldsymbol{U}$ corresponding to small eigenvalues. These are the less-structured functions. In other words, as long as matrix $\boldsymbol{Z}$ is chosen generously enough one is faced with a bias

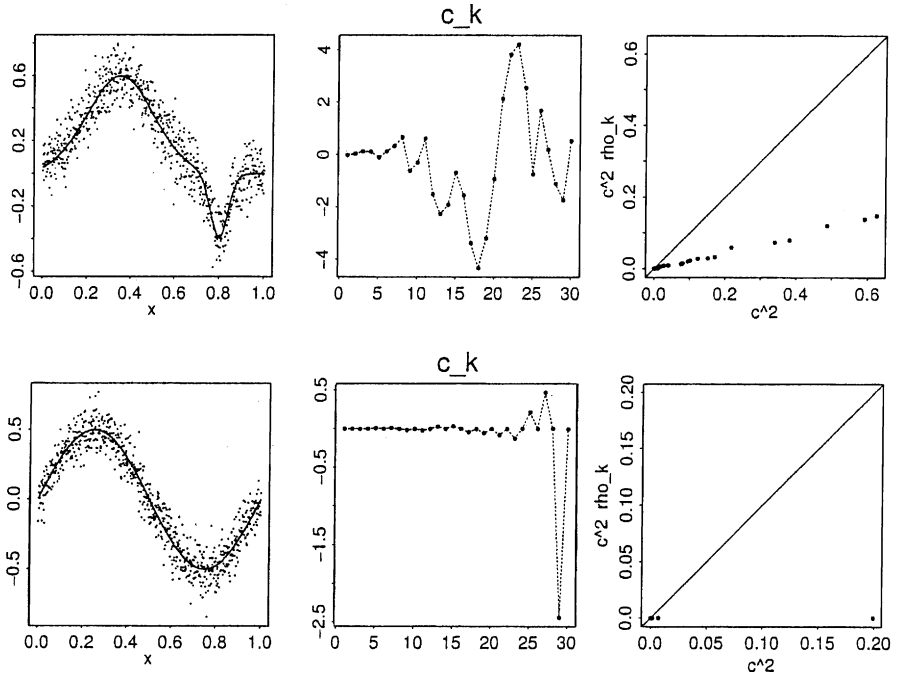

Fig. 2. Coefficients c_k (middle column) for different functions (left column) with a typical sample of size $n = 750$. The right column shows c_k^2/K plotted against $c_k^2\rho_k/\sum_k \rho_k$.

leading to undersmoothing of $\hat{\lambda}_{\text{REML}}$. One should keep in mind that this statement is again formulated in an asymptotic sense and small sample behavior can look differently as the next section will show.

3.2.1. Simulation study

We run a simulation study to investigate the large sample performance in practice. The study will be continued in the next section using a small and moderate sample size. We simulate data from the two functions shown in Fig. 2. The first is $f_1(x) = 1.5\phi\{(x-0.35)/0.15\} - \phi\{(x-0.8)/0.04\}$ with $\phi(\cdot)$ as standard normal density (upper row in Fig. 2) and the second is $f_2(x) = 0.5\sin(2\pi x)$ (bottom row). For fitting we use a $K = 30$ dimensional truncated linear basis. We draw $n = 750$ observations and calculate $\hat{\lambda}_{\text{REML}}$ and $\hat{\lambda}_{\text{Cp}}$ using a 50 dimensional grid search. The corresponding estimated degrees of freedom based on 300 simulations are shown in Fig. 3 in the right hand column. The dotted vertical and horizontal line indicate the optimal MSE choice. The tendency of undersmoothing for $\hat{\lambda}_{\text{REML}}$ is obvious, in particular for the second example. It appears however that for small sample sizes (two right hand columns) the effect looks different for the first function. An explanation for this phenomena will be given in the next section.

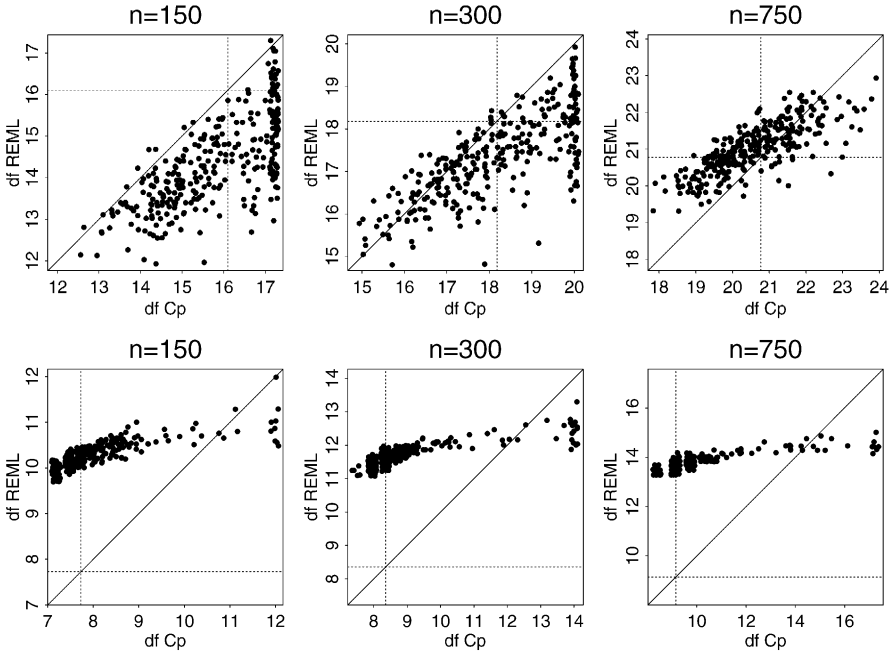

Fig. 3. Selected degrees of freedom $\mathrm{df}(\hat{\lambda}_{\mathrm{Cp}})$ plotted against $\mathrm{df}(\hat{\lambda}_{\mathrm{REML}})$ for different sample sizes. Upper row is for function $f_1(x)$, bottom row for $f_2(x)$. Vertical and horizontal lines show optimal MSE choice.

3.2.2. *Variance*

It can also be observed from the simulation that $\hat{\lambda}_{\mathrm{Cp}}$ is more variable than $\hat{\lambda}_{\mathrm{REML}}$. This can also be shown asymptotically since

$$\mathrm{Var}(\hat{\lambda}_{\mathrm{REML}}) = \frac{2\sigma_\varepsilon^2}{K^2}\sum_{k=1}^{K}\rho_k^2,$$

$$\mathrm{Var}(\hat{\lambda}_{\mathrm{Cp}}) = \frac{2\sigma_\varepsilon^2}{(\sum_{l=1}^{K}\rho_l)^2}\sum_{k=1}^{K}\rho_k^4,$$

which easily proves

$$\mathrm{Var}(\hat{\lambda}_{\mathrm{Cp}}) \geqslant \mathrm{Var}(\hat{\lambda}_{\mathrm{REML}}).$$

4. Finite sample comparison

We will now investigate the rate of convergence in more depth. The results so far are derived up to an asymptotic correction of order $\mathrm{O}(n^{-1})$. It is however well known that asymptotic convergence for smoothing parameter selection criteria may be slow in practice (see Härdle et al., 1988) and it seems therefore worthwhile to explore small

sample properties as well. To do so we first show that the Cp criterion as well as the REML criterion can be comprehended as a penalty concept, where the goodness of fit is penalized by the complexity of the model. This means we write both criteria in the form

$$C(\lambda) = (\boldsymbol{Y} - \hat{\boldsymbol{f}})^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{f}}) + D(\lambda) + \text{const} \tag{18}$$

with $\hat{\boldsymbol{f}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}}$ and const collecting all components not depending on λ. The function $D(\lambda)$ can be comprehended as a measure for complexity of the model. For $\hat{\lambda}_{\mathrm{Cp}}$ we get with (7)

$$D_{\mathrm{Cp}}(\lambda) = 2\sigma_\varepsilon^2 \operatorname{tr}(\boldsymbol{S}_\lambda) = 2\sigma_\varepsilon^2 \operatorname{tr}(\boldsymbol{F}_{Z,X}^{-1}\boldsymbol{F}_{Z,X,\lambda}) + \text{const}$$

with

$$\boldsymbol{F}_{Z,X,\lambda} = n((\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} + \boldsymbol{D}_{K/\lambda}) - \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z})^{-1}.$$

For the REML estimate a decomposition like (18) is less obvious. Using (24) and (25) in the appendix we can however rewrite $C_{\mathrm{REML}}(\lambda) = -\sigma_\varepsilon^2 l_{\mathrm{REML}}(\hat{\boldsymbol{\beta}}, \lambda)$ to

$$\begin{aligned} C_{\mathrm{REML}}(\lambda) &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \sigma_\varepsilon^2 \log|\boldsymbol{V}_\lambda| + \sigma_\varepsilon^2 \log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_\lambda^{-1}\boldsymbol{X}| \\ &= (\boldsymbol{Y} - \hat{\boldsymbol{f}})^{\mathrm{T}}(\boldsymbol{Y} - \hat{\boldsymbol{f}}) + D_{\mathrm{REML}}(\lambda) \end{aligned} \tag{19}$$

with stochastic complexity

$$D_{\mathrm{REML}}(\lambda) = \frac{\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\hat{\boldsymbol{b}}}{\lambda} + K\log(\lambda) - \log|\boldsymbol{F}_{Z,X,\lambda}|. \tag{20}$$

For simplicity of investigation we again assume $\boldsymbol{D}_K = \boldsymbol{I}_K$. Denoting as above with ρ_k the kth eigenvalue of $\boldsymbol{F}_{Z,X}$ we find $\rho_k\{1 - \rho_k/(\rho_k + \lambda n)\}$ as eigenvalue of $\boldsymbol{F}_{Z,X,\lambda}$. This allows us to rewrite the complexities to

$$D_{\mathrm{Cp}}(\lambda) = 2\sigma_\varepsilon^2 \sum_k (\lambda n/(\rho_k + \lambda n)), \tag{21}$$

$$D_{\mathrm{REML}}(\lambda) = \frac{\hat{\boldsymbol{c}}^{\mathrm{T}}\hat{\boldsymbol{c}}}{\lambda} + \sigma_\varepsilon^2 \sum_k \log(\rho_k + \lambda n) \tag{22}$$

with $\hat{\boldsymbol{c}} = \boldsymbol{U}^{\mathrm{T}}\hat{\boldsymbol{b}}$. The objective is now to compare (21) with (22). A conspicuous property of $D_{\mathrm{REML}}(\lambda)$ is that it is non-monotonic. This non-monotonicity implies that small values of λ achieve a large complexity and hence are not selected by the REML criteria. This in fact mirrors the bias towards undersmoothing as the simulation study below will show. Moreover $D_{\mathrm{REML}}(\lambda)$ is stochastic while $D_{\mathrm{MSE}}(\lambda)$ is deterministic. We find asymptotically

$$\hat{\boldsymbol{c}}|\boldsymbol{c} \sim \mathrm{N}\left(\operatorname{diag}(1 - \rho_k/(\rho_k + \lambda n))\boldsymbol{c}, \frac{\sigma_\varepsilon^2}{n}\operatorname{diag}\{\rho_k(1 - \rho_k/(\rho_k + \lambda n))^2\}\right). \tag{23}$$

Note that in (23) we explicitly include terms of order $\mathrm{O}(n^{-1})$ which have been omitted in the previous section. With (23) we can write the stochastic component in $D_{\mathrm{REML}}(\lambda)$ as a weighted sum of non-central Chi-squared distributed variables. This means we get $\hat{\boldsymbol{c}}^{\mathrm{T}}\hat{\boldsymbol{c}} = \sum_k v_k^2 \chi^2_{k,\delta_k}/n$ where $v_k = \sigma_\varepsilon\sqrt{\rho_k}(1 - \rho_k/(\rho_k + \lambda n))$ and χ^2_{k,δ_k} as noncentral

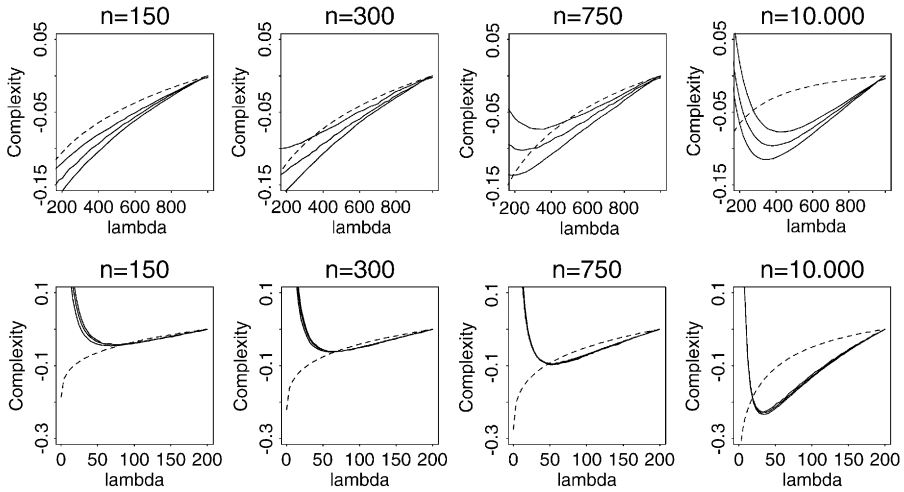

Fig. 4. Complexity measure $D_{\mathrm{REML}}(\lambda)$ with pointwise confidence intervals (solid lines) and $D_{\mathrm{MSE}}(\lambda)$ (dotted lines). Upper row is for function $f_1(x)$, bottom row for $f_2(x)$.

Chi-squared variables, i.e. $\chi^2_{k,\delta_k} = z_k^2$ with $z_k \sim \mathrm{N}(\mu_k, 1)$. The non-centrality parameters are independent of λ and defined through $\mu_k = \sqrt{n}c_k/(\sigma_\varepsilon\sqrt{\rho_k})$. Considering the complexity in more depth we observe that only components μ_k, $k = 1, \ldots, K$, depend on the unknown underlying function and hence small sample behavior of the REML estimate is determined by μ_k exclusively.

4.1. Simulation

We extend the simulation study from the previous section but use small sample sizes of order $n = 150$ and 300. In Fig. 3, we show the resulting estimated degrees of freedom for the functions seen in Fig. 2 (left column). It appears that there is clear undersmoothing taking place, even for small samples for the sinus shape function $f_2(x)$ (bottom row). However for the first function $f_1(x)$ for small n the effect is vice versa and $\hat{\lambda}_{\mathrm{REML}}$ tends to oversmooth. We explore the different behavior for the two functions by plotting the complexity measures $D(\lambda)$. In Fig. 4, we plot $D_{\mathrm{REML}}(\lambda)$ and $D_{\mathrm{MSE}}(\lambda)$ for the three different sample sizes. Additionally we include a plot for a very large sample size $n = 10.000$. For $D_{\mathrm{REML}}(\lambda)$ we include pointwise 95% confidence intervals based on (23). There are various things noticeable from these plots. First and most apparent $D_{\mathrm{REML}}(\lambda)$ is not monotonic. This means in particular that small values of λ exhibit a large complexity when measured with $D_{\mathrm{REML}}(\lambda)$ and hence the routine tends to leave small values of λ unselected. The U shape of $D_{\mathrm{REML}}(\lambda)$ also contributes to the low variance of $\hat{\lambda}_{\mathrm{REML}}$, since small as well as large values of λ are strongly penalized.

From Fig. 4 we also get insight in the different small sample behavior. Considering $D_{\mathrm{REML}}(\lambda)$ for $f_1(x)$ for sample size $n = 150$ (upper left plot) we see that $D_{\mathrm{REML}}(\lambda)$ and $D_{\mathrm{MSE}}(\lambda)$ have a rather similar shape and in fact $D_{\mathrm{REML}}(\lambda)$ shows a larger slope

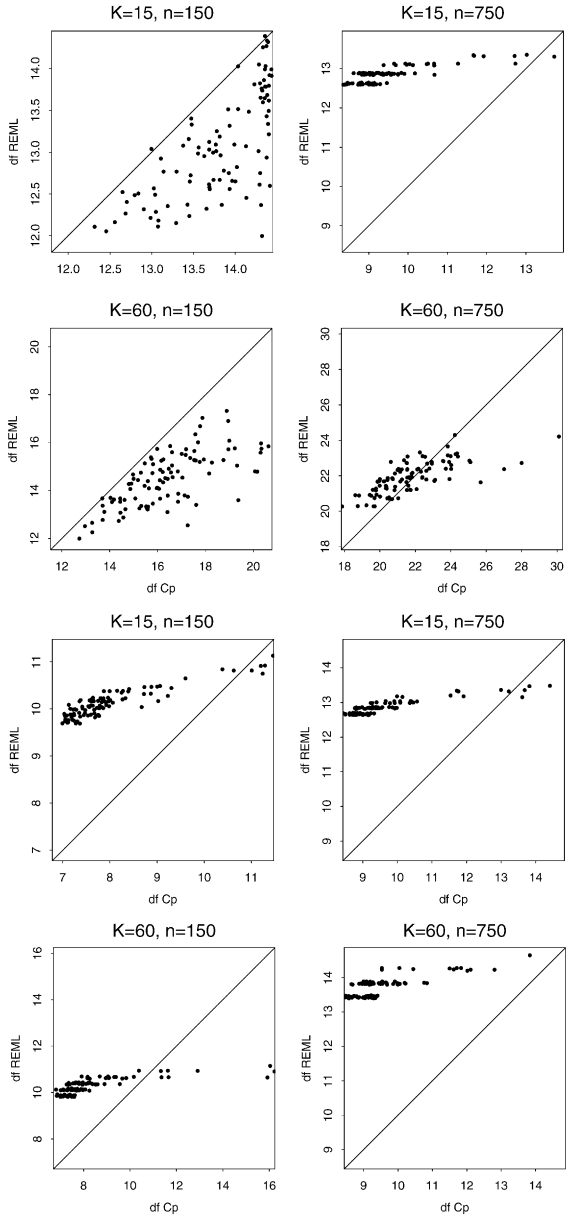

Fig. 5. Selected degrees of freedom for different values of K and n. Upper four plots show results for function $f_1(x)$, lower 4 plots are for function $f_2(x)$.

so that small values of λ are preferred. Accordingly $\hat{\lambda}_{\mathrm{REML}}$ tends to oversmooth as can be seen in Fig. 3 upper left plot. If n increases however the U shape of $D_{\mathrm{REML}}(\lambda)$ becomes dominant and $\hat{\lambda}_{\mathrm{REML}}$ starts undersmoothing. Finally, the minimum of the U

shape of $D_{\mathrm{REML}}(\lambda)$ is larger than the mean of $\hat{\lambda}_{\mathrm{MSE}}$ which expresses the general bias derived in (17).

We extend the simulation to explore the effect of the choice of K. For sample sizes $n=150$ and 750 we build matrix Z from truncated linear lines $(x-\tau_k)_+$, $k=1,\ldots,K$, with τ_k equidistant points on $[0,1]$. We choose $K=15$ and 60 and run for each setting 150 simulations. The results are shown in Fig. 5. For function $f_1(x)$ (upper 4 plots) we observe the same behavior for the larger basis with $K=60$ as in Fig. 3. For $K=15$ the basis seems to be too small so that even for $n=750$ the REML estimate oversmooths. For function $f_2(x)$ undersmoothing of the REML estimate is evident for all settings.

Finally we run some simulations with σ_ε^2 being estimated by $\hat{\sigma}_\varepsilon^2=(\boldsymbol{Y}-\hat{\boldsymbol{f}})^{\mathrm{T}}(\boldsymbol{Y}-\hat{\boldsymbol{f}})/(N-df)$, with df as degree of freedom calculated from the trace of the smoothing matrix. Except of an increased variability of the smoothing parameter estimates the results were the same as those seen in Fig. 3. For space reasons we therefore do not include the resulting plots here.

5. Discussion

In this paper, we compared smoothing parameter selection for P-spline smoothing based on Mean-Squared Error minimization and REML estimation. We discussed different scenarios and showed that the REML estimate has the tendency to undersmooth, i.e. it chooses a too complex model. For small samples this effect can be vice versa depending on the underlying function. The asymptotic result is in line with standard spline smoothing, but the asymptotic scenario is different. While for standard spline fitting the basis grows with the sample size for P-spline smoothing the dimension of the basis is kept fixed. There has been little discussion in the P-spline literature whether K should be fixed independently of n and kept fixed even if n increases. Ruppert (2002) suggests a data based choice of K but also shows that K depends only very little on n.

The problem tackled in this paper was on global smoothing parameter selection. If the function fitted has in fact varying complexity over x a local choice of the smoothing parameter might be more appropriate. This has been suggested in Ruppert and Carroll (2000).

Finally, in concordance with findings in classical spline smoothing the REML estimate shows a reduced variability compared to the Cp alternative. This can be explained asymptotically as well for small samples by the functional form of the criteria.

Appendix A. Technical details

Before deriving asymptotic results we point out the following relationship which is used throughout the paper. Simple matrix algebra shows that

$$\boldsymbol{V}_\lambda^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{Z}\hat{\boldsymbol{b}} \tag{A.1}$$

with $\hat{\boldsymbol{b}}$ as defined in (5). Moreover we get again with simple matrix algebra

$$\begin{aligned}
&\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{Z}\hat{\boldsymbol{b}})\\
&=\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})-\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\boldsymbol{D}_K/\lambda)^{-1}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\\
&=\{\boldsymbol{I}-\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}+\boldsymbol{D}_K/\lambda)^{-1}\}\boldsymbol{Z}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})\\
&=\boldsymbol{D}_K\hat{\boldsymbol{b}}/\lambda. \qquad \text{(A.2)}
\end{aligned}$$

A.1. MSE smoothing parameter

We assume the conditional model shown in the second component of (9) with $\boldsymbol{b}$ unknown but fixed. Covariate $\boldsymbol{x}$ is assumed to have compact support with density bounded away from zero. Denoting $\boldsymbol{W}=(\boldsymbol{X},\boldsymbol{Z})$ we postulate that $\boldsymbol{F}_W=n(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}$ is a matrix of order O(1). The estimate $\hat{\boldsymbol{f}}_\lambda=\{\hat{f}(x_1),\dots\hat{f}(x_n)\}^{\mathrm{T}}$ is obtained by $\hat{\boldsymbol{f}}=\boldsymbol{S}_\lambda\boldsymbol{Y}$ with

$$\boldsymbol{S}_\lambda=\boldsymbol{W}\{\boldsymbol{F}_W^{-1}+\boldsymbol{D}/(\lambda n)\}^{-1}\boldsymbol{W}^{\mathrm{T}}$$

as smoothing matrix, where $\boldsymbol{D}$ as block diagonal matrix $\mathrm{diag}(\boldsymbol{0}_p,\boldsymbol{D}_K)$, where $\boldsymbol{0}_p$ is a matrix of zeros with p as number of columns in $\boldsymbol{X}$. Subsequently we take advantage of expansions of the type

$$\{\boldsymbol{F}_W^{-1}+\boldsymbol{D}/(\lambda n)\}^{-1}=\boldsymbol{F}_W-\frac{1}{\lambda n}\boldsymbol{F}_W\boldsymbol{D}\boldsymbol{F}_W+\frac{1}{(\lambda n)^2}\boldsymbol{F}_W\boldsymbol{D}\boldsymbol{F}_W\boldsymbol{D}\boldsymbol{F}_W+\cdots. \qquad \text{(A.3)}$$

This allows for the bias $\boldsymbol{B}=E(\hat{\boldsymbol{f}})-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b}$ the approximation

$$\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=\left\{\frac{1}{\lambda^2 n}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\boldsymbol{b}-\frac{2}{\lambda^3 n^2}\boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\boldsymbol{b}\right\}\{1+\mathrm{O}(n^{-1})\}$$

and for the variance we get the decomposition

$$\mathrm{tr}\left\{\mathrm{Var}(\hat{\boldsymbol{f}})\right\}=\sigma_\varepsilon^2\left\{(p+K)-\frac{2}{\lambda n}\mathrm{tr}(\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K)+\frac{3}{(\lambda n)^{-2}}\mathrm{tr}(\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K)+\cdots\right\}.$$

Differentiating $\mathrm{MSE}(\lambda)=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}+\mathrm{tr}\{\mathrm{Var}(\hat{\boldsymbol{f}})\}$ leads to optimal MSE estimate given in (6).

A.2. CP estimate

Using approximation arguments as above it is easy to see that

$$\frac{\partial}{\partial\lambda}\mathrm{tr}(\boldsymbol{S}_\lambda)=\frac{1}{\lambda^2 n}\mathrm{tr}(\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K)-\frac{2}{\lambda^3 n^2}\mathrm{tr}(\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K)+\cdots.$$

Moreover, straightforward calculation shows

$$\frac{\partial}{\partial\lambda}(\boldsymbol{Y}-\hat{\boldsymbol{f}})^{\mathrm{T}}(\boldsymbol{Y}-\hat{\boldsymbol{f}})=-\frac{1}{\lambda^3 n}\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\hat{\boldsymbol{b}}+\frac{1}{\lambda^4 n^2}\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\boldsymbol{F}_{Z\!X}\boldsymbol{D}_K\hat{\boldsymbol{b}}+\cdots.$$

Employing this to set the derivative of (7) to zero directly proves (8).

A.3. REML estimate

Differentiation of (12) with respect to λ yields

$$\frac{\partial l_{\mathrm{REML}}(\boldsymbol{\beta},\lambda)}{\partial\lambda}$$

$$=\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}_{\lambda}^{-1}\boldsymbol{Z}\boldsymbol{D}_K^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\lambda}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{\sigma_{\varepsilon}^2} \tag{A.4}$$

$$-\operatorname{tr}(\boldsymbol{V}_{\lambda}^{-1}\boldsymbol{Z}\boldsymbol{D}_K^{-1}\boldsymbol{Z}^{\mathrm{T}})+\operatorname{tr}\{(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{V}_{\lambda}^{-1}\boldsymbol{X})(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_{\lambda}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_{\lambda}^{-1}\boldsymbol{Z})\}. \tag{A.5}$$

Replacing $\boldsymbol{\beta}$ in (A.4) by its estimate (4) allows to simplify (A.4) to $\hat{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{D}_K\hat{\boldsymbol{b}}/\lambda^2$. Moreover, simple matrix manipulation similar to (A.3) provides to expand (A.5) which gives the leading components $-K/\lambda+\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)/(\lambda^2 n)$. This in turn proves (13).

A.4. Comparison

We assume now that model (9) holds. Considering estimate $\hat{\boldsymbol{b}}\equiv\hat{\boldsymbol{b}}_n$ conditional on $\boldsymbol{b}$ provides with simple asymptotic arguments

$$\hat{\boldsymbol{b}}_n|\boldsymbol{b}\overset{a}{\sim}\mathrm{N}\left\{(\boldsymbol{I}_K-\frac{1}{\lambda n}\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)\boldsymbol{b}+\mathrm{O}(n^{-2}),\sigma_{\varepsilon}^2\boldsymbol{F}_{Z.X}/n+\mathrm{O}(n^{-2})\right\}, \tag{A.6}$$

so that with (9) the joint probability results as

$$\begin{pmatrix}\boldsymbol{b}\\ \hat{\boldsymbol{b}}\end{pmatrix}\overset{a}{\sim}\mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\sigma_{\varepsilon}^2\begin{pmatrix}\lambda\boldsymbol{D}_K^{-1} & \lambda\boldsymbol{D}_K^{-1}-\boldsymbol{F}_{Z.X}/n\\ \lambda\boldsymbol{D}_K^{-1}-\boldsymbol{F}_{Z.X}/n & \lambda\boldsymbol{D}_K^{-1}-\boldsymbol{F}_{Z.X}/n\end{pmatrix}\right)=\mathrm{N}(0,\boldsymbol{\Sigma}_n).$$

Solving the REML equation (A.4) and (A.5) up to the second asymptotic order and defining $z_n=\sigma_{\varepsilon}^2(\hat{\lambda}_{\mathrm{REML}}-\lambda_{\mathrm{MSE}})$ gives

$$\begin{aligned}z_n &= (\boldsymbol{b}^{\mathrm{T}},\hat{\boldsymbol{b}}^{\mathrm{T}})\begin{pmatrix}-\dfrac{\boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K}{\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)} & 0\\ 0 & \dfrac{\boldsymbol{D}_K}{K-\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)/n}\end{pmatrix}\begin{pmatrix}\boldsymbol{b}\\ \hat{\boldsymbol{b}}\end{pmatrix}+\delta/n\\ &=:(\boldsymbol{b}^{\mathrm{T}},\hat{\boldsymbol{b}}^{\mathrm{T}})\boldsymbol{A}\begin{pmatrix}\boldsymbol{b}\\ \hat{\boldsymbol{b}}\end{pmatrix}+\delta/n,\end{aligned}$$

where $\delta=\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)/(\sigma_{\varepsilon}^2K)-3\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)/(\sigma_{\varepsilon}^2\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K))$. The focus is now to calculate $P(z_n\leqslant 0)$. We tackle this problem using an Edgeworth expansion (see e.g. McCullagh, 1987, pp. 147, 148) by approximating the distribution of z_n by $z_{\infty}=\lim_{n\to\infty}z_n$. Since

$$z_{\infty}=\boldsymbol{b}^{\mathrm{T}}\left(-\frac{\boldsymbol{D}_K\boldsymbol{F}_{Z.X}\boldsymbol{D}_K}{\operatorname{tr}(\boldsymbol{F}_{Z.X}\boldsymbol{D}_K)}+\frac{\boldsymbol{D}_K}{K}\right)\boldsymbol{b}$$

we get with Imhof (1961)

$$z_\infty = \sigma_b^2 \sum_{k=1}^{K} (1/K - \rho_k / \sum_l \rho_l) \chi_k^2 \tag{A.7}$$

with χ_k^2 as independent Chi-squared distributed variable with 1 degree of freedom and ρ_k as eigenvalues of $\boldsymbol{F}_{Z,X}\boldsymbol{D}_K$. Analogously we find $z_n = \sum_{k=1}^{2K} \rho_{(n)k} \chi^2_{k,\delta_k} + \delta/n$ where χ_{k,δ_n} are now noncentral independent Chi-squared variables with δ_k as noncentrality parameter and $\rho_{(n)k}$ as characteristic roots of $\boldsymbol{A}\boldsymbol{\Sigma}_n$. It is easy to see that $\rho_{(n)k} = \mathrm{O}(n^{-1})$ for K of the roots while the remaining K roots fulfill $\rho_{(n)k} = \rho_k\{1 + \mathrm{O}(n^{-1})\}$ after appropriate reordering. Moreover we find

$$E(z_n) - E(z_\infty) = \delta/n \tag{A.8}$$

and analogously differences in higher-order cumulants of z_n and z_∞ are of negligible asymptotic order. This allows to write

$$P(z_n \leqslant 0) = P(z_\infty \leqslant 0) - \frac{h(0)\delta}{Kn} + \mathrm{O}(n^{-2}),$$

where $h(\cdot)$ denotes the density of z_∞. With (A.7) we finally obtain (14).

References

Aerts, M., Claeskens, G., Wand, M., 2002. Some theory for penalized spline additive models. J. Statist. Plan. Inference 103, 455–470.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed model. J. Amer. Statist. Assoc. 88, 9–25.

Brumback, B.A., Rice, J.A., 1998. Smoothing spline models for the analysis of nested and crossed samples of curves (c/r: P976-994). J. Amer. Statist. Assoc. 93, 961–976.

Davies, R., 1980. The distribution of a linear combination of χ^2 random variables. Appl. Statist. 29, 323–333.

Efron, B., 2001. Selection criteria for scatterplot smoothers. Ann. Statist. 29, 470–504.

Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statist. Sci. 11 (2), 89–121.

French, J., Kammann, E., Wand, M., 2001. Comment on paper by Ke and Wang. J. Amer. Statist. Assoc. 96, 1285–1288.

Green, D.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London.

Härdle, W., Hall, P., Marron, J.S., 1988. How far are automatically chosen regression smoothing parameters from their optimum? (c/r: P96-101). J. Amer. Statist. Assoc. 83, 86–95.

Harville, D., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc. 72, 320–338.

Imhof, J., 1961. Computing the distribution of quadratic forms in normal variables. Biometrika 48, 419–426.

Kou, S., Efron, B., 2002. Smoothers and the Cp, GML and EE criteria: a geometric approach. J. Amer. Statist. Assoc. 97, 766–782.

Li, K., 1985. From Stein's unbiased risk estimates to the method of generalized crossvalidation. Ann. Statist. 13, 1352–1377.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

McCullagh, P., 1987. Tensor Methods in Statistics. Chapman & Hall, London.

O'Sullivan, F., 1986. A statistical perspective on ill-posed inverse problems (c/r: P519-527). Statist. Sci. 1, 502–518.

Parker, R., Rice, J., 1985. Discussion of "Some aspects of the spline smoothing approach to nonparametric curve fitting" by B.W. Silverman. J. Roy. Statist. Soc. Ser. B 47, 40–42.

Pinheiro, J., Bates, D., 2000. Mixed-Effects Models in *S* and *S*plus. Springer, New York.

Ruppert, D., 2002. Selecting the number of knots for penalized splines. J. Comput. Graphical Statist. 11, 735–757.

Ruppert, D., Carroll, R., 2000. Spatially-adaptive penalties for spline fitting. Austral. New Zealand J. Statist. 42, 205–224.

Searle, S., Casella, G., McCulloch, C., 1992. Variance Components. Wiley, New York.

Speckman, P., Sun, D., 2001. Asymptotic properties of smoothing parameter selection in spline smoothing. Technical Report, Department of Statistics, University of Missouri.

Stein, M.L., 1990. A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. Ann. Statist. 18, 1139–1157.

Verbyla, A.P., Cullis, B.R., Kenward, M.G., Welham, S.J., 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. Appl. Statist. 48, 269–311.

Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Statist. 13, 1378–1402.

Wand, M., 1999. On the optimal amount of smoothing in penalised spline regression. Biometrika 86, 936–940.

Wand, M., 2003. Smoothing and mixed models. Comput. Statist. 18, 223–249.

Wecker, W.E., Ansley, C.F., 1983. The signal extraction approach to nonlinear regression and spline smoothing. J. Amer. Statist. Assoc. 78, 81–89.