Computational Statistics and Data Analysis 53 (2009) 2550–2562

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Generalized profiling estimation for global and adaptive penalized spline smoothing

Jiguo Cao[a,*], James O. Ramsay[b]

[a] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada

[b] Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, QC, H3A 1B1, Canada

ARTICLE INFO

Article history:
Received 16 July 2008
Received in revised form 18 November 2008
Accepted 5 December 2008
Available online 25 December 2008

ABSTRACT

We propose the generalized profiling method to estimate the multiple regression functions in the framework of penalized spline smoothing, where the regression functions and the smoothing parameter are estimated in two nested levels of optimization. The corresponding gradients and Hessian matrices are worked out analytically, using the Implicit Function Theorem if necessary, which leads to fast and stable computation. Our main contribution is developing the modified delta method to estimate the variances of the regression functions, which include the uncertainty of the smoothing parameter estimates. We further develop adaptive penalized spline smoothing to estimate spatially heterogeneous regression functions, where the smoothing parameter is a function that changes along with the curvature of regression functions. The simulations and application show that the generalized profiling method leads to good estimates for the regression functions and their variances.

1. Introduction

Nonparametric regression, or smoothing, describes the flexible association between covariates and responses, and many competing methods have been proposed, including kernel-based methods and spline smoothing. We consider the representation of a sample of N functional observations by a set of smooth curves. Let $t_{ij}, i = 1, \ldots, n_j; j = 1, \ldots, N$ be a point at which the jth process is observed, and let y_{ij} be the observed value. No restrictions on the t_{ij}'s, such as equal spacing or the same values for all observations, are required, although we do assume that the observations are defined over a common domain $t \in \Omega$. For simplicity, we adopt the additive Gaussian error model

$$y_{ij} = \mu_j(t_{ij}|\boldsymbol{\beta}) + \epsilon_{ij},$$

where $\boldsymbol{\epsilon}_j$ is an n_j-vector containing the measurement errors for observation j and is assumed in the distribution of $MN(0, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\Sigma}_j$ is the corresponding variance–covariance matrix. The functional parameters $\mu_j(t|\boldsymbol{\beta})$ may depend on a finite dimensional vector $\boldsymbol{\beta}$ of fixed effect parameters that do not vary with j, as well as on further functional parameters through models such as the semiparametric regression model $\mu_j[\boldsymbol{\beta}'\mathbf{z}_j + \eta_j(t)]$, where $\mathbf{z}_j$ is a vector of known covariates. If there is not such a dependency, we shall use the notation $\mu_j(t)$. The estimator $x_j(t)$ of $\mu_j(t)$ will be estimated with the spline method (Wahba, 1983, 1990; Friedman, 1991; Wand, 2000; de Boor, 2001). In other words, $x_j(t)$ is expressed in terms of the basis function expansion

$$x_j(t) = \mathbf{c}_j'\boldsymbol{\phi}(t)$$

[*] Corresponding author.
E-mail addresses: jca76@sfu.ca (J. Cao), ramsay@psych.mcgill.ca (J.O. Ramsay).

0167-9473/$ – see front matter Crown Copyright © 2008 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.12.004

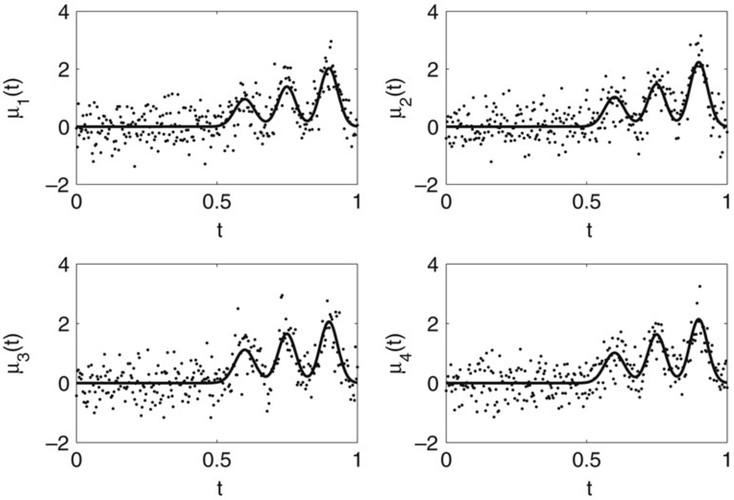

Fig. 1. One set of simulated data (dots) generated by adding Gaussian noise (STD = 0.5) to 300 equally spaced points. The solid lines are the true curve $\mu_i = a_{1i}\exp(-400(x-0.6)^2) + a_{2i}\exp(-500(x-0.75)^2) + a_{3i}\exp(-500(x-0.9)^2)$, $a_{1i} \sim \texttt{Normal}(1, 0.1^2)$, $a_{2i} \sim \texttt{Normal}(5/3, 0.1^2)$, $a_{3i} \sim \texttt{Normal}(2, 0.1^2)$, $i = 1, \ldots, 4$.

where $\boldsymbol{\phi}$ is a vector of length K_μ of known basis functions. There are many good basis systems, such as Fourier basis and wavelets. In particular, it has been proven that any piecewise smooth function can be well approximated by the spline basis, which is defined by a sequence of knots. de Boor (2001) showed how to improve the spline approximation accuracy and efficiency by the knot selection. However, there are few methods that can select the optimal knot sequence for arbitrary problems.

Instead, we prefer to put at least one knot on each point having an observation, so that the basis function expansion is powerful enough to capture any amount of variation in the observed data. To prevent the estimated function from overfitting the data, we require a roughness penalty in our optimization criterion, which is called penalized smoothing. The roughness penalty is defined as

$$\texttt{PEN}(x_j) = \lambda \int_\Omega \|L_{\boldsymbol{\gamma}} x_j(t)\|^2 \mathrm{d}t.$$

This functional is defined by a possibly nonlinear differential operator $L_{\boldsymbol{\gamma}}$ that may in turn depend on a parameter vector $\boldsymbol{\gamma}$ whose value has to be estimated. Norms other than that of $\mathcal{L}_2$ may be used depending on the situation.

All simulations and applications in this article use the second derivative to define the roughness penalty term. Although the paper is developed within the framework of generalized least squares estimation, the methodology that we propose can be applied to more general distributions and estimation procedures such as maximum likelihood or Bayesian estimation. The generalized profiling methodology that we propose for the estimation of λ has already been used by Cao and Ramsay (2007) for the estimation of location parameter $\boldsymbol{\beta}$, and by Ramsay et al. (2007) for the estimation of the operator parameter $\boldsymbol{\gamma}$, and to keep the exposition as simple as possible, we will consider both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, if required, as known. Further background for the problem is available in Ramsay and Silverman (2005).

We extend the penalized spline smoothing to estimate spatially heterogeneous regression functions, where the smoothing parameter $\lambda(t)$ is a function that changes along with the curvature of regression functions. This procedure is called *adaptive penalized smoothing*. Fig. 1 illustrates the type of regression problems where adaptive penalized smoothing may be beneficial. We want to estimate, from noisy observations, multiple functions μ_j, $j = 1, \ldots, N$, that have sharp curvature over one or more small regions and rather mild curvature elsewhere. It seems reasonable to enforce a much lower level of smoothness in the neighborhoods with high curvature. The adaptive smoothing parameter estimate can be seriously improved when a sample of N curves are to be estimated, all having such sharp features in similar regions. A single adaptive smoothing parameter is estimated from observations for the N curves, incorporates the similarities of these curves. The vectors of basis coefficients are different for each curve, capturing the specificity of each curve. Fig. 2 displays the logarithm of an adaptive smoothing parameter found by applying our method to the curves in Fig. 1, which does just what is required.

The adaptive smoothing is implemented by defining the penalty functional $\texttt{PEN}(x_j)$ as

$$\texttt{PEN}(x_j) = \int_\Omega \lambda(t)\|L_{\boldsymbol{\gamma}} x_j(t)\|^2 \mathrm{d}t. \tag{1}$$

The penalty $\texttt{PEN}(x_j)$ is defined by a possibly nonconstant weight function $\lambda(t) > 0$. Permitting λ to be a function rather the usual practice of regarding λ as a constant allows the possibility that a stronger penalty may be appropriate for some regions

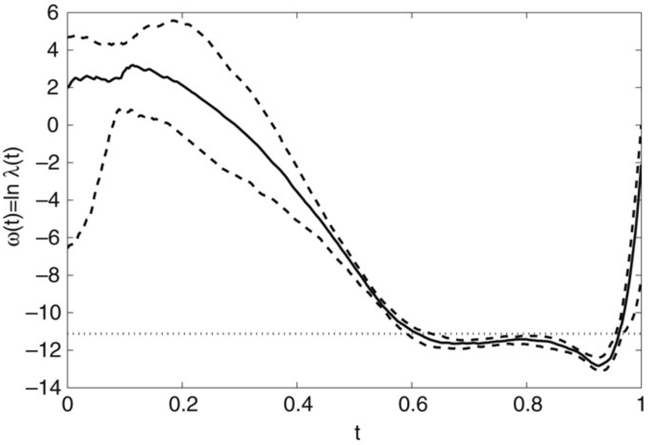

Fig. 2. The functional smoothing parameter estimates changes with the spatial heterogeneity of curves shown in Fig. 1. The solid line is the median of the estimated functional smoothing parameter $\omega(t) = \ln\lambda(t)$. The dashed lines are the 25% and 75% quantiles of the estimated functional smoothing parameter $\omega(t)$. The dotted straight line is the estimated constant $\omega = \ln\lambda$ in global penalized smoothing.

in Ω where the behavior of μ_j is close to being in the null space or kernel of $L_{\boldsymbol{\gamma}}$ than for other regions. For example, if $L = D^2$ (the classic choice defining cubic spline smoothing), it may be that μ_j is nearly linear over some regions, where its estimate can be heavily penalized relative to penalties applied over other regions where it displays a strong second derivative.

The primary contribution of our paper is to obtain confidence regions for estimates of μ_j that take into account the uncertainty passed along to these estimates by any data-driven approach to choose the amount of smoothing. The usual practice of ignoring this uncertainty in estimating confidence regions (Gu, 2002) becomes problematic when the estimated smoothing parameters, such as that shown in Fig. 2, require multiple parameters to define.

There is a considerable literature on adaptive smoothing within the contexts of kernel and local polynomial smoothing methods (Hardle and Bowman, 1983; Staniswalis, 1989; Friedman and Silverman, 1989; Vieu, 1991; Brockmann et al., 1993; Eubank and Speckman, 1993; Fan and Gijbels, 1995; Fan et al., 1996; Lepski et al., 1997; Herrmann, 1997), among others). Boularan et al. (1995) and Nunez-Anton et al. (1999) took into account a possible common structure of a family of curves and estimated these curves with nonparametric kernel smoothing techniques. To apply adaptive smoothing techniques to spline estimates, Nychka (1995) linked the spline method with the kernel smoothing by showing in theories that the absolute value of the spline weight function decreased exponentially away from its center. Ruppert and Carroll (2000) considered the spatial heterogeneity of the regression function and proposed to penalize the P-spline coefficients adaptively. They estimated the standard error of the regression function with the empirical Bayesian method, which ignored the uncertainty of the smoothing parameter estimate. Baladandayuthapani et al. (2005) constructed a Bayesian version of this local penalty method, and estimated the regression function and the smoothing parameter simultaneously. Their method can require intensive computation, and it may be hard to implement for the Naive users without the Bayesian Background.

Our adaptive penalized spline smoothing method has three unique aspects. First, the smoothing parameter is a function, which can be adapted to the spatial heterogeneity of the data. Second, we estimate the regression curves and the functional smoothing parameter in two nested levels of optimization. This approach is able to converge more easily than the simultaneous estimation approach, and the computation is also faster than the Bayesian method. Finally, the variance estimate for the regression curves includes the uncertainty in the estimate of the functional smoothing parameter.

Our paper is organized as follows. Section 2 introduces the generalized profiling method for the nuisance and complexity parameter estimates. Section 3 introduces the modified delta method applied to estimate the standard errors for parameter estimates, which include the uncertainty of other parameter estimates. Section 4 compares our modified delta method and the empirical Bayesian method in the standard error estimates for the regression function by simulation. The adaptive penalized spline smoothing method is also compared with the global penalized spline smoothing method, the global kernel smoothing method and the local kernel smoothing method. Section 5 illustrates the adaptive penalized smoothing method with an application. Section 6 contains discussion and conclusions.

2. Generalized profiling estimation for parameters

Our general setup in the previous section specifies three classes of parameters that require estimation from the $n = \sum_j n_j$ scalar data values. The N coefficient arrays $\mathbf{c}_j$ have the usual characteristics of *nuisance* parameters. First of all, their number tends to be design-dependent in the sense that the more data that one has per curve, the larger the number K_μ of basis functions that one is likely to use, and this is especially the case if the common practice in using spline bases of placing a knot at each value t_{ij} is followed. Secondly, the actual value of any specific coefficient c_{jk} is seldom of great interest; rather the coefficients are required to model a type of variation in the data that cannot be ignored. Moreover, it may be desirable to consider them as randomly sampled in a multilevel sampling structure. Finally, the sheer number of coefficients to be estimated is typically far larger than the parameters in the next two classes that we now consider.

The parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, on the other hand, *structural* parameters of the classic sort; they are of fixed and limited dimensionality, are often the primary focus in the data analysis, and are typically viewed as fixed in multilevel designs. The functional smoothing parameter $\lambda(t)$ is estimated with the expansion after the logarithm transformation:

$$\lambda(t) = \exp[\omega(t)], \quad \text{where } \omega(t) = \boldsymbol{\theta}'\boldsymbol{\psi}(t),$$

where $\boldsymbol{\psi}(t)$ is a vector of K_ω basis functions. The parameter vector $\boldsymbol{\theta}$ who defines $\lambda(t)$ seems to have a status of its own. First of all, like structural parameters, it will be only in a small dimension; but, like the coefficients, it may also have some design dependency because large amounts of data per curve and large numbers of curves will probably tempt users to employ a richer basis. In fact, $\boldsymbol{\theta}$ has some of the character of random effects variance parameters in multilevel designs, and for this reason we call it the *model complexity* parameter.

The central idea behind the generalized profiling method or the parameter cascade notion described in Cao and Ramsay (2007) and Ramsay et al. (2007) is to treat these three parameter classes in quite different ways through the use of a *multicriterion optimization* approach. We describe this approach in this section, where we assume that the variance–covariance matrices $\boldsymbol{\Sigma}_j, j = 1, \dots, N$, are known. As we have indicated above, in order to focus on the role of the complexity parameters $\boldsymbol{\theta}$, we will further simplify our situation by assuming that the structural parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are either known or are not present in the problem at hand.

Conditional on current values for $\boldsymbol{\theta}$, and consistent with the assumption of multinormally distributed residual vectors $\boldsymbol{\epsilon}_j$, we employ in the analyses reported in this paper the criterion

$$H(\mathbf{c}_j|\boldsymbol{\theta}, \mathbf{y}_j) = \|\mathbf{y}_j - \mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)\|^2 + \int \lambda(t)[Lx(t|\mathbf{c}_j)]^2 \mathrm{d}t, \tag{2}$$

where vector $\mathbf{y}_j$ contains the n_j values y_{ij} and $\mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)$ contains the corresponding values $x(t_{ij}|\mathbf{c}_j)$, and where

$$\|\mathbf{y}_j - \mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)\|^2 = [\mathbf{y}_j - \mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)]'\mathbf{W}_j[\mathbf{y}_j - \mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)]$$

for some known order n_j weighting matrix $\mathbf{W}_j$ such as $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1}$. More generally, however, the loss function in the first term of (2) may be a negative log likelihood or any other suitable measure of lack of fit.

Define order K_μ matrix

$$\mathbf{R} = \int \lambda(t)[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'\mathrm{d}t,$$

and let $\boldsymbol{\Phi}_j$ be the $n_j \times K_\mu$ matrix with the ikth element $\phi_k(t_{ij})$. By minimizing $H(\mathbf{c}_j|\lambda(t), \mathbf{y}_j)$, we can estimate the coefficient vector $\mathbf{c}_j$, and in the particular case (2) the optimal value written analytically as

$$\hat{\mathbf{c}}_j(\lambda(t), \mathbf{y}_j) = [\boldsymbol{\Phi}_j'\mathbf{W}_j\boldsymbol{\Phi}_j + \mathbf{R}]^{-1}\boldsymbol{\Phi}_j'\mathbf{W}_j\mathbf{y}_j. \tag{3}$$

When $\int[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'\mathrm{d}t$ is an identity matrix, and $\lambda(t)$ is a step function defined as $\lambda(t) = \lambda_k$ when $t \in \{t : \phi_k(t) > 0\}$, $k = 1, \dots, K_\mu$, then the criterion (2) reduces to

$$H^*(\mathbf{c}_j|\lambda(t), \mathbf{y}_j) = \|\mathbf{y}_j - \mathbf{x}(\mathbf{t}_j|\mathbf{c}_j)\|^2 + \sum_{k=1}^{K_\mu} \lambda_k c_{jk}^2,$$

which is the adaptive smoothing criterion proposed by Ruppert and Carroll (2000).

The optimal functional smoothing parameter $\hat{\lambda}(t)$ can be chosen by minimizing a complexity measure such as the generalized cross-validation (GCV, Wahba, 1985)

$$\mathtt{GCV}(\lambda(t)) = \left[\frac{n}{\mathtt{dfe}(\lambda(t))}\right]\left[\frac{\mathtt{SSE}(\lambda(t))}{\mathtt{dfe}(\lambda(t))}\right],$$

where $n = \sum_j n_j$, both the degrees of freedom measure $\mathtt{dfe}(\lambda(t))$, and the sum of squared errors $\mathtt{SSE}(\lambda(t))$ can be written in terms of the order n_j matrix $\mathbf{A}_j(\lambda(t)) = \boldsymbol{\Phi}_j(\boldsymbol{\Phi}_j'\mathbf{W}_j\boldsymbol{\Phi}_j + \mathbf{R})^{-1}\boldsymbol{\Phi}_j'\mathbf{W}_j$:

$$\mathtt{dfe}(\lambda(t)) = n - \sum_{j=1}^{N} \mathtt{Tr}[\mathbf{A}_j(\lambda(t))];$$

$$\mathtt{SSE}(\lambda(t)) = \sum_{j=1}^{N} \{\mathbf{y}_j'[I - \mathbf{A}(\lambda(t))]'[I - \mathbf{A}(\lambda(t))]\mathbf{y}_j\}.$$

Note that $\mathtt{GCV}(\lambda(t))$ only depends on $\lambda(t)$; $\hat{\mathbf{c}}(\lambda(t)) = (\hat{\mathbf{c}}_1(\lambda(t))', \dots, \hat{\mathbf{c}}_N(\lambda(t))')'$ has disappeared from the problem because the conditional solution (3) defines it as a function of $\lambda(t)$. We may call $\hat{\mathbf{c}}[\lambda(t)]$ an *estimating function*, and its role is to act as a pipeline or conduit that channels the fitting power defined by a $\lambda(t)$-function into local fitting parameters $\hat{c}_{jk}[\lambda(t)]$, which is the kth element of $\hat{\mathbf{c}}_j[\lambda(t)]$. That is, $\lambda(t)$ is like a water reservoir that is used to irrigate a crop, and $\hat{\mathbf{c}}(\lambda(t))$ is the irrigation system that transports the water to various areas in the field, where $\hat{c}_{jk}(\lambda(t))$ sprinkles the growing seedlings, namely the $\mathbf{y}_j$'s, in sphere of influence of each basis function k. If our basis were of the Fourier type, then the locations would be defined in terms of frequency, if wavelets, in terms of locations on the frequency/time plane, and so on.

In the adaptive spline smoothing framework, there are two kinds of parameters: the nuisance parameter $\mathbf{c}_j$ and the complexity parameter $\boldsymbol{\theta}$. They are estimated in two nested levels of optimization. In the inner optimization level where we minimize $H(\mathbf{c}_j|\boldsymbol{\theta}, \mathbf{y}_j)$, the nuisance parameter $\mathbf{c}_j$ is estimated conditional on $\boldsymbol{\theta}$ and $\mathbf{y}_j$, and thus the conditional estimate $\hat{\mathbf{c}}_j$ can be treated as a function of $\boldsymbol{\theta}$ and $\mathbf{y}_j$. In the outer optimization level, denote $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})$ to be the outer optimization criterion, which is GCV in this article. The nuisance parameter $\hat{\mathbf{c}}$ is removed from the parameter space as a function of $\boldsymbol{\theta}$ and $\mathbf{y}$. The complexity parameter $\boldsymbol{\theta}$ is then estimated by minimizing $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})$.

The functional relationship between the nuisance and complexity parameters allows us to calculate the gradient and Hessian matrix analytically in both levels of optimization, which is essential for fast computation. Denote $\mathrm{d}F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})/\mathrm{d}\boldsymbol{\theta}$ to be the total derivative of $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})$ with respect to $\boldsymbol{\theta}$, then we have

$$\frac{\mathrm{d}F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})}{\mathrm{d}\boldsymbol{\theta}} = \frac{\partial F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})}{\partial \boldsymbol{\theta}} + \frac{\partial F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})}{\partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}.$$

The derivative $\partial \hat{\mathbf{c}}/\partial \boldsymbol{\theta}$ is required in the above formula, and it is also crucial for the variance estimation of parameters. When $\hat{\mathbf{c}}$ is an implicit function of $\boldsymbol{\theta}$, the Implicit Function Theorem is applied as follows. The estimate $\hat{\mathbf{c}}$ satisfies $\partial H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})/\partial \mathbf{c} = 0$, and consequently

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \left(\left. \frac{\partial H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}} \right|_{\hat{\mathbf{c}}} \right) = \left. \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta}} \right|_{\hat{\mathbf{c}}} + \left. \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2} \right|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} = 0. \tag{4}$$

If it can be assumed that $\left|\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})/\partial \mathbf{c}^2|_{\hat{\mathbf{c}}}\right| \neq 0$, then we have

$$\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} = -\left[\left. \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2} \right|_{\hat{\mathbf{c}}} \right]^{-1} \left[\left. \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta}} \right|_{\hat{\mathbf{c}}} \right].$$

3. Interval estimation for parameters

By considering the functional relationship between the nuisance and structural parameters, we can obtain their unconditional variance estimates by the Delta method.

Let $\boldsymbol{\mu}$ be the expectation of $\mathbf{y}$, then we can approximate $\hat{\boldsymbol{\theta}}$ as a function of $\mathbf{y}$ with the first order Taylor expansion:

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) \approx \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) + \left[\left. \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} \right|_{\mathbf{y}=\boldsymbol{\mu}} \right] (\mathbf{y} - \boldsymbol{\mu}). \tag{5}$$

Taking variance on both sides of (5), we can approximate the variance of $\hat{\boldsymbol{\theta}}$:

$$\mathtt{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \approx \left[\left. \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} \right|_{\mathbf{y}=\boldsymbol{\mu}} \right] \boldsymbol{\Sigma} \left[\left. \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} \right|_{\mathbf{y}=\boldsymbol{\mu}} \right]',$$

where $\boldsymbol{\Sigma}$ is the variance–covariance matrix for $\mathbf{y}$. Wahba (1983) estimates it by:

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathtt{SSE}(\hat{\boldsymbol{\theta}})}{\mathtt{dfe}(\hat{\boldsymbol{\theta}})} \cdot \mathbf{I}.$$

The derivative of $\mathrm{d}\hat{\boldsymbol{\theta}}/\mathrm{d}\mathbf{y}$ can also be calculated by the Implicit Function Theorem. The estimate $\hat{\boldsymbol{\theta}}$ satisfies $\mathrm{d}F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})/\mathrm{d}\boldsymbol{\theta} = 0$. We take the $\mathbf{y}$-derivative on both sides of the identity $\mathrm{d}F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})/\mathrm{d}\boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} = 0$, and attain:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}} \left(\left. \frac{\mathrm{d}F}{\mathrm{d}\boldsymbol{\theta}} \right|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} \right) = \left. \frac{\mathrm{d}^2 F}{\mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{y}} \right|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} + \left. \frac{\mathrm{d}^2 F}{\mathrm{d}\boldsymbol{\theta}^2} \right|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} \frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} = 0, \tag{6}$$

where

$$\frac{\mathrm{d}^2 F}{\mathrm{d}\boldsymbol{\theta}^2} = \frac{\partial^2 F}{\partial \boldsymbol{\theta}^2} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}} \partial \boldsymbol{\theta}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} \right)' \frac{\partial^2 F}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2}, \tag{7}$$

and

$$\frac{\mathrm{d}^2 F}{\mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{y}} = \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}} \partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \mathbf{y}}. \tag{8}$$

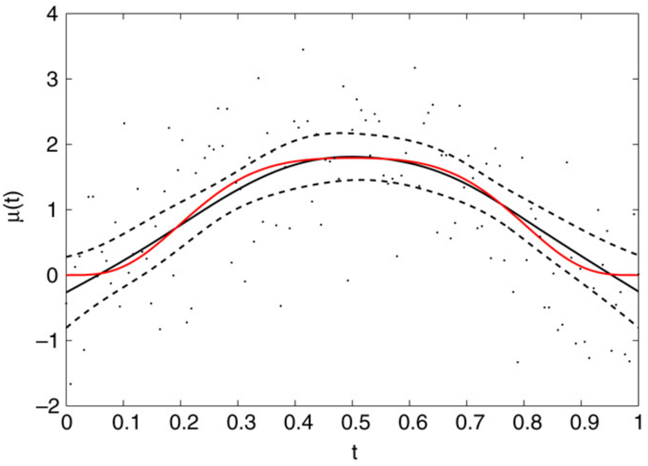

Fig. 3. The global penalized smoothing curves (black solid lines) with the 95% confidence intervals (black dashed lines). The red solid lines indicate the true curves $\mu(t) = 1/3\beta_{10,5}(t) + 1/3\beta_{7,7}(t) + 1/3\beta_{5,10}(t)$, $t \in [0, 1]$. The simulation is explained in detail in Section 4.1. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Note that the final terms in (7) and (8) are products of a vector and a three-way array or tensor with the inner product being taken across the vectors' index. The calculations for $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$, $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}^2$ and $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial\mathbf{y}$ are given in Appendix A.

The first derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$ is acquired by solving (6):

$$\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} = -\left[\left.\frac{\mathrm{d}^2F}{\mathrm{d}\boldsymbol{\theta}^2}\right|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\right]^{-1}\left[\left.\frac{\mathrm{d}^2F}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}}\right|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\right].$$

Similarly, the sampling variance of $\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})$ is

$$\mathtt{Var}[\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})] \approx \left[\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right]',$$

where

$$\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}} = \frac{\partial\hat{\mathbf{c}}}{\partial\hat{\boldsymbol{\theta}}}\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}.$$

If we do not consider the functional relationship between $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\theta}}$, we will obtain the *conditional sampling variance* for $\hat{\mathbf{c}}$

$$\mathtt{Var}[\hat{\mathbf{c}}|\hat{\boldsymbol{\theta}}, \mathbf{y}] \approx \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\boldsymbol{\Sigma}\left(\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\right)'. \tag{9}$$

The conditional sampling variance $\mathtt{Var}[\hat{\mathbf{c}}|\hat{\boldsymbol{\theta}}, \mathbf{y}]$ underestimates the variance of $\hat{\mathbf{c}}$, because it ignores the uncertainty of $\hat{\boldsymbol{\theta}}$. The pointwise estimate for the standard error of $x(t)$ is given by

$$\hat{\sigma}_x^2(t) = \boldsymbol{\phi}'(t)\left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)\hat{\boldsymbol{\Sigma}}\left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)'\boldsymbol{\phi}(t).$$

4. Simulations

4.1. A simulated global penalized smoothing problem

One function in the example of Wahba (1983), which is $\mu(t) = 1/3\beta_{10,5}(t) + 1/3\beta_{7,7}(t) + 1/3\beta_{5,10}(t)$, $t \in [0, 1]$, is used to generate 1000 simulated data sets by adding Gaussian noise with a standard deviation of 1 to 129 equally spaced points in [0, 1]. Here $\beta_{a,b}(t)$ indicates the Beta probability density function with parameters a and b. Fig. 3 displays a typical set of noisy data. The estimate for $\omega = \ln(\lambda)$ is -4.0 with 95% confidence interval $[-6.8, -1.2]$. Although the data have a large scale of noise, the regression function estimated with the global penalized smoothing is close to the true function. Fig. 3 also shows the point confidence interval for the regression functions, which is calculated by $[\hat{\mu}(t) - 1.96 * \hat{\sigma}_\mu(t), \hat{\mu}(t) + 1.96 * \hat{\sigma}_\mu(t)]$. We define the coverage probability as the percentage of the values of the true function at 129 points covered by the pointwise confidence intervals, averaged over 1000 simulations. The coverage probability is 94.6% when the confidence interval calculated with the generalized profiling method, while the empirical Bayesian method only leads to 91.5% coverage probability. This is because our standard error estimate $\hat{\sigma}_\mu(t)$ includes the uncertainty of the smoothing parameter estimates.

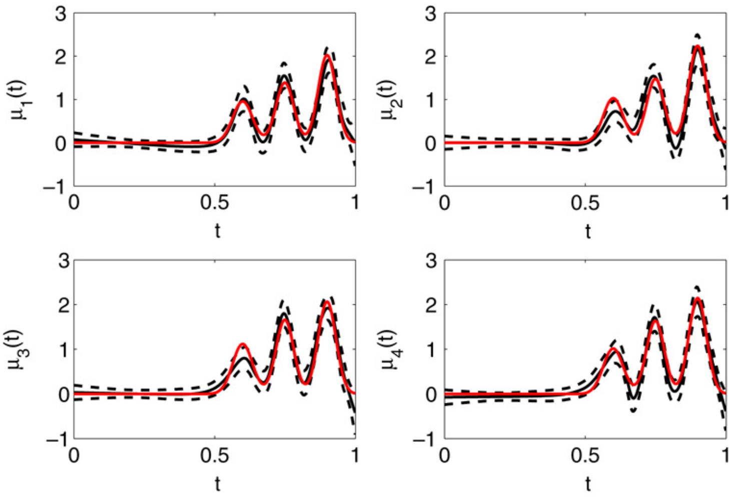

Fig. 4. The adaptive penalized smoothing curves (black solid lines) with the 95% confidence intervals (black dashed lines). The red solid lines indicate the true curves. The simulation is explained in detail in Section 4.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

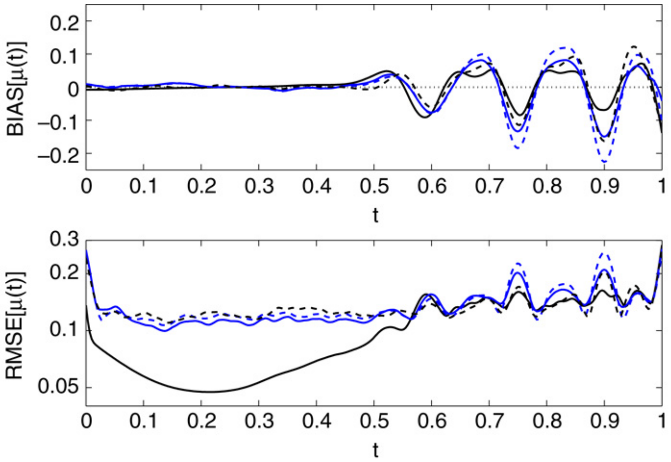

Fig. 5. The bias and root mean squared error of the adaptive penalized smoothing curves, which are averaged over four curves. The black solid line corresponds to adaptive penalized spline smoothing, the black dashed lines correspond to global spline penalized smoothing, the blue dashed lines correspond to global kernel smoothing, and the blue solid lines correspond to local kernel smoothing. The simulation is explained in detail in Section 4.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. *A simulated adaptive smoothing problem*

Four spatially heterogeneous curves are generated with forms $\mu_i = a_{1i}\exp(-400(t-0.6)^2)+a_{2i}\exp(-500(t-0.75)^2)+a_{3i}\exp(-500(t-0.9)^2)$, where $a_{1i} \sim \texttt{Normal}(1, 0.1^2)$, $a_{2i} \sim \texttt{Normal}(5/3, 0.1^2)$, $a_{3i} \sim \texttt{Normal}(2, 0.1^2)$, $i = 1, \ldots, 4$, $t \in [0, 1]$. These functions are suspected to be better estimated with the adaptive penalized smoothing approach, since they show different level of smoothness in the whole region, flat when $t < 0.5$, but rough when $t \geq 0.5$. 1000 simulated data sets are generated by adding Gaussian noise with a standard deviation of 0.5 to 300 equally spaced points drawn from each of the four functions.

Fig. 1 displays a typical set of simulated observations along with the true curves. The regression functions are expanded with cubic B-splines with 40 equally spaced knots. We report results for the smoothing parameter $\omega(t)$ defined as a constant and as a cubic B-spline basis expansion with interior knots placed at 0.5, 0.6, 0.75 and 0.9. Fig. 2 shows that the estimated functional smoothing parameter $\hat{\omega}(t) = \ln(\hat{\lambda}(t))$ ranges from its lowest value of about -13 in [0.5, 1] to 3 in [0 0.5). The traditional penalized smoothing, corresponding to $\hat{\omega} = \ln(\hat{\lambda}) = -11$, by comparison, over-smooths slightly on the right side but under-smooths drastically on the left side. Fig. 4 shows that the adaptive penalized smoothing provides a good estimate of $\mu_i(t)$, even in the region of high curvature.

The adaptive penalized spline smoothing method is compared with the global penalized spline smoothing method, the global kernel smoothing method and the local kernel smoothing method. Fig. 5 displays the pointwise bias, root mean

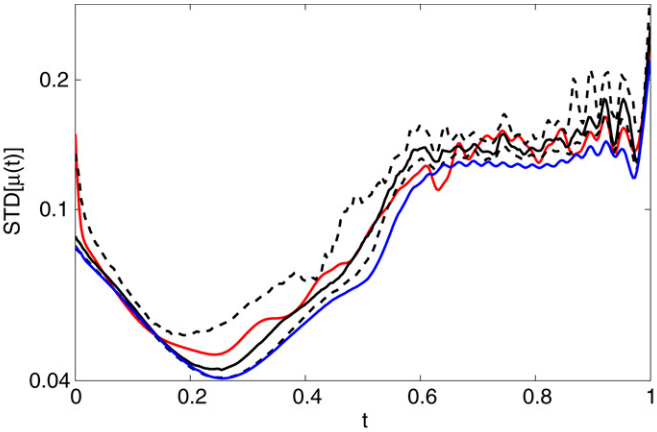

Fig. 6. The median of standard error estimates for the first curve $\mu_1(t)$ over 100 simulations. The black solid line indicates the median of standard error estimates from the modified delta methods. The blue solid line indicates the median of standard error estimates from the empirical Bayesian method (Ruppert and Carroll, 2000). The red solid line indicates the empirical standard deviation of the adaptive penalized smoothing curves. The black dashed lines indicates the 25% and 75% quantiles of standard error estimates from the modified delta methods. The simulation is explained in detail in Section 4.2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

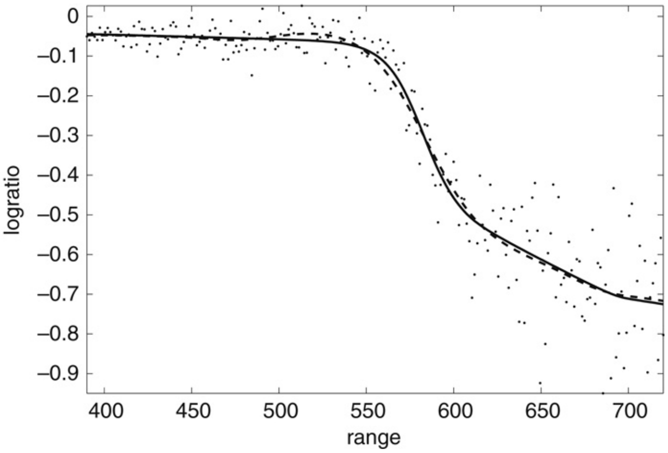

Fig. 7. The horizontal variable, range, is the distance of an object to the laser source. The vertical variable, logratio, is the logarithm of the ratio of received-signal frequencies on and off the resonance frequency of mercury. The solid and dashed lines are the regression functions estimated by adaptive and non-adaptive penalized smoothing, respectively.

squared error (RMSE) of the curve estimates using the four methods over 1000 simulations. The adaptive penalized spline smoothing method leads much smaller pointwise RMSE on the left side and slightly smaller pointwise bias on the right side than the other three methods.

The estimated standard error $\hat{\sigma}_x(t)$ of the curve shown in the top left panel of Fig. 1 is shown in Fig. 6, where we see that the empirical standard deviation for the fitted curves is close to the median of the estimated standard errors. Fig. 6 also shows that the median of standard error estimates from the empirical Bayesian method (Ruppert and Carroll, 2000) is smaller than the empirical standard deviation for the fitted curves because they ignore the uncertainty of the estimates for the functional smoothing parameter. The same conclusion are obtained for the estimated standard errors of other three curves in Fig. 1.

5. Application

The LIDAR (Light Detection and Ranging) is an optical remote sensing technology which measures the distance of an object by measuring the time delay between sending a laser-emitted light and detecting the reflected signal (Sigrist, 1994). LIDAR technology has a wide application in geology, atmospheric physics, and a host of other areas.

Fig. 7 displays a typical LIDAR data set, which was taken from Holst et al. (1996) and Ruppert and Carroll (2000). The range is the distance of an object to the laser source, and the ordinate is the logarithm of the ratio of received-signal frequencies on and off the resonance frequency of mercury. The regression function is expanded with cubic B-splines with one knot placed

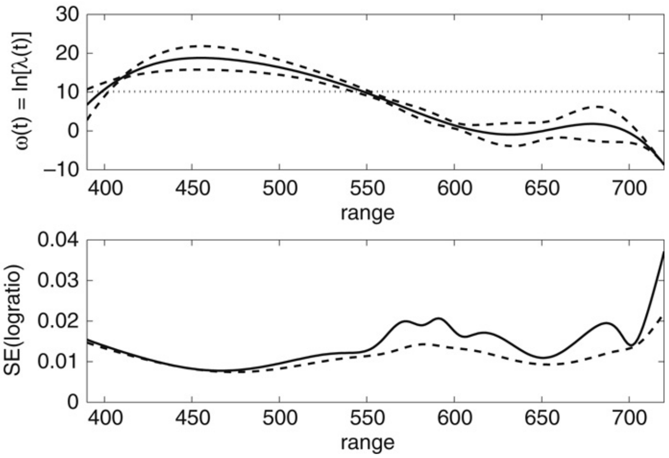

Fig. 8. The top panel shows the optimal adaptive smoothing parameter (the solid line) with its 95% confidence interval (dashed lines). The optimal constant smoothing parameter is shown as the dotted line. The bottom panel displays the standard errors for the adaptive estimate of the regression function when we include or ignore the uncertainty of the estimate for the adaptive smoothing parameter, which are plotted with solid and dashed lines, respectively.

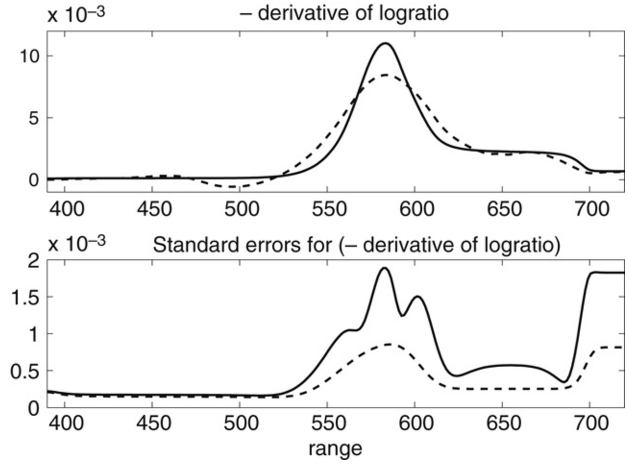

Fig. 9. The top panel shows the derivatives of the regression functions estimated with adaptive smoothing (solid line) and non-adaptive smoothing (dashed line), respectively. The bottom panel displays the standard errors for the derivatives of the adaptive regression function when including or ignoring the uncertainty of the estimate for the adaptive smoothing parameter, which are plotted with solid and dashed lines, respectively.

on the ith quantile of the range, $i = 0$, 1%, 2%, . . . , 100%. The functional smoothing parameter $\omega(t) = \ln(\lambda(t))$ is expanded using cubic B-splines with 4 interior knots on the 50%, 60%, 75% and 99% quantiles of the range. We do not put interior knots in the left side of the range, since the logratio has a flat trend in that interval. The solid and dashed lines are the regression functions estimated by adaptive and non-adaptive penalized smoothing respectively. The two lines are close to each other, except that the non-adaptive regression function has a small bump in the range [500, 550].

The top panel in Fig. 8 shows the optimal adaptive smoothing parameter with its 95% confidence interval. Compared with the optimal constant smoothing parameter (the dotted line), the functional smoothing parameter is larger in the left side and smaller in the right side. So we have a larger penalty in the left side for the regression function, and consequently we obtain a more smooth regression function with the adaptive penalized smoothing, which correctly reflects the LIDAR data. The bottom panel in Fig. 8 displays the standard error of the regression function using adaptive smoothing, which includes the uncertainty of the adaptive smoothing parameter estimates and is larger than that obtained with the empirical Bayesian method (Ruppert and Carroll, 2000).

It is also of scientific interest to estimate the first derivative of the regression function, $x'(t)$ (Ruppert et al., 1997). The top panel in Fig. 9 displays the derivatives of the regression functions estimated with adaptive smoothing, which has a sharper peak and less variability in the flat areas than that obtained with non-adaptive smoothing. It is also positive in the whole interval of the range, which is meaningful as a concentration function. The standard error for the adaptive regression function is shown in the bottom panel of Fig. 9, which is also larger than that obtained with the empirical Bayesian method (Ruppert and Carroll, 2000), which ignores the uncertainty of adaptive smoothing parameter estimates.

6. Discussion and conclusions

The idea of treating the complexity controller λ, or, preferably, $\omega = \ln \lambda$, as a parameter to be estimated from the data is hardly new (Gu, 2002), and the usual method of optimizing the GCV measure with $\mathbf{c}$ re-estimated with each new trial λ value is essentially the profiling process that we have outlined. However, by considering the structural parameter space to be unidimensional in the constant λ situation, we have been able to develop useful interval estimates for both λ and the coefficient vector $\mathbf{c}(\lambda)$ on which it depends. The latter estimates are superior to the usual conditional estimate in taking account of the uncertainty in the λ estimate that defines them. In our simulation results, the coverage of these new interval estimates seems quite reasonable.

Within this framework, it is natural to consider λ to be a higher dimensional parameter, and we see that adaptive penalized smoothing defined in this way can bring important benefits by applying less smoothing where there is more curvature, and more where the curvature is minimal.

The adaptive penalized smoothing problem that we have considered here is somewhat special in that $\lambda(t)$ controls model complexity, and consequently it is critical to use an outer optimization criterion that measures the complexity of $x(t)$ in a way that is mathematically independent of the total lack of fit. The GCV criterion as usually defined can no doubt be improved in this regard when adaptive penalized smoothing is used, and this is a subject for further research.

The penalized smoothing problem involves both the nuisance and structural parameters. We propose the generalized profiling method to estimate these two different types of parameters in two levels of optimization, allowing a different criterion in each level. We consider the functional relationship of nuisance and structural parameters, which is the key for decreasing the computation load and finding the unconditional variance estimates. We also want to propose this approach to the treatment of nuisance parameters in a much wider context. Problems of this nature are found everywhere in statistics, and include, for example, multilevel linear models and psychometric models.

All of the results in this paper have been generated by programming in the Matlab computing language, making use of functional data analysis software intended to compliment Ramsay and Silverman (2005). A general function for adaptive smoothing with an example is available from the URL: http://www.stat.sfu.ca/~cao/Research.html.

Acknowledgements

This paper has benefitted substantially from the constructive comments of an associate editor and two referees. This research was supported by grants from the Natural Science and Engineering Research Council of Canada to both authors. The authors thank Dr. David Ruppert for contributing the LIDAR data analyzed in the article.

Appendix A. Derivative calculations for estimating variances of global and local parameters

The formulas (7) and (8) for $\mathrm{d}^2F/\mathrm{d}\boldsymbol{\theta}^2$ and $\mathrm{d}^2F/\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{y}$ involve the terms $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$, $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}^2$ and $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial\mathbf{y}$. In the following, we derive the formulas for these three terms.

We introduce the following convention, which is called *Einstein Summation Notation*. If a Latin index is repeated in a term, then it is understood as a summation with respect to that index. For instance, instead of the expression $\sum_i a_i x_i$, we merely write $a_i x_i$.

- $\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$

  Since the optimal nuisance parameter vector $\hat{\mathbf{c}}$ satisfying $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial\mathbf{c} = 0$, and $\hat{\mathbf{c}}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$, we can take the $\mathbf{y}$-derivative on $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial\mathbf{c}|_{\hat{\mathbf{c}}} = 0$ as follows:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{y}}\left(\left.\frac{\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}}\right|_{\hat{\mathbf{c}}}\right) = \left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\mathbf{y}}\right|_{\hat{\mathbf{c}}} + \left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} = 0, \tag{A.1}$$

which holds since $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial\mathbf{c}|_{\hat{\mathbf{c}}}$ is a function of $\mathbf{y}$ that is identically 0. Assuming that $\left|\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial\mathbf{c}^2|_{\hat{\mathbf{c}}}\right| \neq 0$, from the Implicit Function Theorem we obtain

$$\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} = -\left[\left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\right]^{-1}\left[\left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\mathbf{y}}\right|_{\hat{\mathbf{c}}}\right]. \tag{A.2}$$

- $\frac{\partial\hat{\mathbf{c}}^2}{\partial\boldsymbol{\theta}\partial\mathbf{y}}$

  We take the y_k-derivative on both sides of Eq. (4):

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}dy_k}\left(\left.\frac{\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}}\right|_{\hat{\mathbf{c}}}\right) &= \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\right|_{\hat{\mathbf{c}}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k} \\ &\quad + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial y_k}\right|_{\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial\hat{c}_i}{\partial y_k}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial y_k} \\ &= 0. \end{aligned} \tag{A.3}$$

Solving for $\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial y_k}$, we obtain the second derivative of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\theta}$ and y_k:

$$\begin{aligned}\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial y_k} = -\left[\left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\right]^{-1} &\left[\left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\right|_{\hat{\mathbf{c}}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k}\right. \\ &\left. + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial y_k}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right] \end{aligned} \tag{A.4}$$

- $\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2}$

  Similar to (A.4), the second partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ and θ_j is:

$$\begin{aligned}\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \theta_j} = -\left[\left.\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\right]^{-1} &\left[\left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial \theta_j}\right|_{\hat{\mathbf{c}}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial \theta_j}\right. \\ &\left. + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial \theta_j}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \left.\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial \theta_j}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right]. \end{aligned} \tag{A.5}$$

Appendix B. Matrix calculations for adaptive penalized smoothing

We provide here the results required for estimates of pointwise standard errors of the complexity function $\omega(t)$ in adaptive penalized smoothing (Section 4). In order to simplify notation, we define the order K_c matrix $\mathbf{B}(\lambda) = \boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R}$ and order n matrix $\mathbf{A}(\lambda) = \boldsymbol{\Phi}\mathbf{B}(\lambda)^{-1}\boldsymbol{\Phi}'\mathbf{W}$. To make the mathematical formula more readable, we assume the number to curve $N = 1$. Then we can express $\mathtt{SSE}(\lambda)$ and degrees of freedom measure $\mathtt{dfe}(\lambda)$ in terms of the matrix A:

$$\begin{aligned}&\mathtt{SSE}(\lambda) = \mathbf{y}'[I - \mathbf{A}(\lambda)]'[I - \mathbf{A}(\lambda)]\mathbf{y} \\ &\mathtt{dfe}(\lambda) = n - \mathrm{trace}(\mathbf{A}(\lambda)).\end{aligned}$$

In what follows, we suppress the explicit dependence of these three matrices on λ and the parameter vector $\boldsymbol{\theta}$ in order to keep the notation readable.

- The first derivatives with respect to the $\omega(t)$ basis coefficient θ_l of these three matrices are:

$$\begin{aligned}&\frac{\partial \mathbf{R}}{\partial \theta_l} = \int \lambda(t)\psi_l(t)[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'\mathrm{d}t \\ &\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l} = -\mathbf{B}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_l}\mathbf{B}^{-1} \\ &\frac{\partial \mathbf{A}}{\partial \theta_l} = \boldsymbol{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l}\boldsymbol{\Phi}'\mathbf{W}.\end{aligned}$$

- The second derivatives with respect to the smoothing function basis coefficients θ_l and θ_i are:

$$\begin{aligned}&\frac{\partial^2 \mathbf{R}}{\partial \theta_l \partial \theta_i} = \int \lambda(t)\psi_i(t)\psi_l(t)[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'\mathrm{d}t \\ &\frac{\partial^2 \mathbf{B}^{-1}}{\partial \theta_l \partial \theta_i} = -\frac{\partial \mathbf{B}^{-1}}{\partial \theta_i}\frac{\partial \mathbf{R}}{\partial \theta_l}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial^2 \mathbf{R}}{\partial \theta_l \partial \theta_i}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_l}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_i} \\ &\frac{\partial^2 \mathbf{A}}{\partial \theta_l \partial \theta_i} = \boldsymbol{\Phi}\frac{\partial^2 \mathbf{B}^{-1}}{\partial \theta_l \partial \theta_i}\boldsymbol{\Phi}'\mathbf{W}.\end{aligned}$$

- The first derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficient θ_l is

$$\frac{\partial \mathtt{GCV}(\lambda)}{\partial \theta_l} = n\left[\mathtt{dfe}\frac{\partial \mathtt{SSE}}{\partial \theta_l} - 2\mathtt{SSE}\frac{\partial \mathtt{dfe}}{\partial \theta_l}\right]\mathtt{dfe}^{-3} \tag{B.1}$$

  where

$$\begin{aligned}&\frac{\partial \mathtt{dfe}(\lambda)}{\partial \theta_l} = -\mathrm{trace}\left(\frac{\partial \mathbf{A}}{\partial \theta_l}\right) \\ &\frac{\partial \mathtt{SSE}(\lambda)}{\partial \theta_l} = -\mathbf{y}'\left(\left[\frac{\partial \mathbf{A}}{\partial \theta_l}\right]'[\mathbf{I} - \mathbf{A}] + [\mathbf{I} - \mathbf{A}]'\left[\frac{\partial \mathbf{A}}{\partial \theta_l}\right]\right)\mathbf{y}.\end{aligned}$$

- The second derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficients θ_l and θ_j is

$$\begin{aligned}\frac{\partial^2\mathtt{GCV}(\lambda)}{\partial\theta_l\partial\theta_j} &= \frac{n}{\mathtt{dfe}^2}\frac{\partial^2\mathtt{SSE}}{\partial\theta_l\partial\theta_j} - \frac{2n\mathtt{SSE}}{\mathtt{dfe}^3}\frac{\partial^2\mathtt{dfe}}{\partial\theta_l\partial\theta_j} + \frac{6n\mathtt{SSE}}{\mathtt{dfe}^4}\frac{\partial\mathtt{dfe}}{\partial\theta_l}\frac{\partial\mathtt{dfe}}{\partial\theta_j} \\ &\quad - \frac{2n}{\mathtt{dfe}^3}\left[\frac{\partial\mathtt{dfe}}{\partial\theta_l}\frac{\partial\mathtt{SSE}}{\partial\theta_j} + \frac{\partial\mathtt{dfe}}{\partial\theta_j}\frac{\partial\mathtt{SSE}}{\partial\theta_l}\right]\end{aligned} \tag{B.2}$$

where

$$\frac{\partial^2\mathtt{SSE}(\lambda)}{\partial\theta_l\partial\theta_j} = \mathbf{y}'(E'+E)\mathbf{y}$$

$$\frac{\partial^2\mathtt{dfe}(\lambda)}{\partial\theta_l\partial\theta_j} = -\text{trace}\left(\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial\theta_j}\right)$$

and

$$E = \left[\frac{\partial\mathbf{A}}{\partial\theta_l}\right]'\left[\frac{\partial\mathbf{A}}{\partial\theta_j}\right] - \left[\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial\theta_j}\right]'[\mathbf{I}-\mathbf{A}].$$

- The second derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficients θ_l and $\mathbf{y}$ is

$$\frac{\partial^2\mathtt{GCV}(\lambda)}{\partial\theta_l\partial\mathbf{y}} = n\left[\mathtt{dfe}\frac{\partial^2\mathtt{SSE}}{\partial\theta_l\partial\mathbf{y}} - 2\frac{\partial\mathtt{SSE}}{\partial\mathbf{y}}\frac{\partial\mathtt{dfe}}{\partial\theta_l}\right]\mathtt{dfe}^{-3} \tag{B.3}$$

where

$$\frac{\partial\mathtt{SSE}(\lambda)}{\partial\mathbf{y}} = 2[\mathbf{I}-\mathbf{A}]'[\mathbf{I}-\mathbf{A}]\mathbf{y}$$

$$\frac{\partial^2\mathtt{SSE}(\lambda)}{\partial\theta_l\partial\mathbf{y}} = -2\left\{\left[\frac{\partial\mathbf{A}}{\partial\theta_l}\right]'[\mathbf{I}-\mathbf{A}] + [\mathbf{I}-\mathbf{A}]'\frac{\partial\mathbf{A}}{\partial\theta_l}\right\}\mathbf{y}.$$

- The sampling variance of $\omega(t) = \ln\lambda(t)$ is estimated by:

$$\mathtt{Var}(\omega(t)) = \left(\frac{\mathrm{d}\omega}{\mathrm{d}\mathbf{y}}\right)\boldsymbol{\Sigma}\left(\frac{\mathrm{d}\omega}{\mathrm{d}\mathbf{y}}\right)' \tag{B.4}$$

where

$$\frac{\mathrm{d}\omega}{\mathrm{d}\mathbf{y}} = \boldsymbol{\psi}(t)'\left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\mathbf{y}}\right) \quad \text{and} \quad \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}\mathbf{y}} = \left[\frac{\partial^2\mathtt{GCV}(\lambda)}{\partial^2\boldsymbol{\theta}}\right]^{-1}\frac{\partial^2\mathtt{GCV}(\lambda)}{\partial\boldsymbol{\theta}\partial\mathbf{y}}.$$

- Since the estimated curve $\hat{\mathbf{x}}(t) = \hat{\mathbf{c}}'\boldsymbol{\phi}(t)$, we can estimate the sampling variance of $\hat{\mathbf{x}}(t)$ by

$$\mathtt{Var}[\hat{\mathbf{x}}(t)] = \boldsymbol{\phi}'(t)\mathtt{Var}(\hat{\mathbf{c}})\boldsymbol{\phi}(t). \tag{B.5}$$

where

$$\mathtt{Var}[\hat{\mathbf{c}}] = \left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)\boldsymbol{\Sigma}\left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)',$$

$$\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}} = \mathbf{B}^{-1}\boldsymbol{\Phi}'\mathbf{W} + \sum_l \frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\theta_l}\frac{\mathrm{d}\theta_l}{\mathrm{d}\mathbf{y}},$$

$$\text{and} \quad \frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\theta_l} = \frac{\mathrm{d}\mathbf{B}^{-1}}{\mathrm{d}\theta_l}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y}.$$

References

Baladandayuthapani, V., Mallick, B.K., Carroll, R.J., 2005. Spatially adaptive Bayesian penalized regression splines (P-splines). Journal of Computational and Graphical Statistics 14, 378–394.

Boularan, J., Ferre, L., Vieu, P., 1995. A nonparametric model for unbalanced longitudinal data with application to geophysical-data. Computational Statistics 10 (3), 285–298.

Brockmann, M., Gasser, T., Herrmann, E., 1993. Locally adaptive bandwidth choice for kernel regression estimators. Journal of the American Statistical Association 88 (424), 1302–1309.

Cao, J., Ramsay, J.O., 2007. Parameter cascades and profiling in functional data analysis. Computational Statistics 22 (3), 335–351.

de Boor, C., 2001. A Practical Guide to Splines. Springer, New York.

Eubank, R.L., Speckman, P., 1993. Confidence bands in nonparametric regression. Journal of the American Statistical Association 88, 1287–1301.

Fan, J., Gijbels, I., 1995. Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. Journal of Royal Statistical Society. Section B 57 (2), 371–394.

Fan, J., Hall, P., Martin, M.A., Patil, P., 1996. On local smoothing of nonparametric curve estimators. Journal of the American Statistical Association 91, 258–266.

Friedman, J.H., 1991. Multivariate adaptive regression splines (with discussion). The Annals of Statistics 19, 1–141.

Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling (with discussion). Technometrics 31, 3–39.

Gu, C., 2002. Smoothing Spline Anova Models. Springer, New York.

Hardle, W., Bowman, A., 1983. Bootstrapping in nonparametric regression – local adaptive smoothing and confidence bands. Journal of the American Statistical Association 83, 102–110.

Herrmann, E., 1997. Local bandwidth choice in kernel regression estimation. Journal of Computational and Graphical Statistics 6 (1), 35–54.

Holst, U., Hüssjer, O., Bjürklund, C., Ragnarson, P., Edner, H., 1996. Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. Environmetrics 7, 401–416.

Lepski, O.V., Mammen, E., Spokoiny, V.G., 1997. Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. The Annals of Statistics 25 (3), 929–947.

Nunez-Anton, V., Rodriguez-Poo, J.M., Vieu, P., 1999. Longitudinal data with nonstationary errors: A nonparametric three-stage approach. Test 8, 201–231.

Nychka, D., 1995. Splines as local smoothers. The Annals of Statistics 23, 1175–1197.

Ramsay, J.O., Hooker, G., Campbell, D., Cao, J., 2007. Parameter estimation for differential equations: A generalized smoothing approach (with discussion). Journal of the Royal Statistical Society. Series B 69, 741–796.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. Springer, New York.

Ruppert, D., Carroll, R.J., 2000. Spatially-adaptive penalties for spline fitting. Australian and New Zealand Journal of Statistics 42 (2), 205–223.

Ruppert, D., Wand, M.P., Holst, U., Hössjer, O., 1997. Local polynomial variance function estimation. Technometrics 39, 262–273.

Sigrist, M., 1994. Air Monitoring by Spectroscopic Technique. In: Chemical Analysis Series, vol. 127. Wiley.

Staniswalis, J.G., 1989. Local bandwidth selection for kernel estimates. Journal of the American Statistical Association 84, 284–288.

Vieu, P., 1991. Nonparametric regression: Optimal local bandwidth choice. Journal of the Royal Statistical Society. Series B 53, 453–464.

Wahba, G., 1983. Bayesian confidence intervals for the cross-validated smoothing spline. Journal of the Royal Statistical Society. Series B 45, 133–150.

Wahba, G., 1985. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. The Annals of Statistics 13 (4), 1378–1402.

Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia.

Wand, M.P., 2000. A comparison of regression spline smoothing procedures. Computational Statistics 15 (4), 443–462.