Computational Statistics (2007) 22:335–351
DOI 10.1007/s00180-007-0044-1

ORIGINAL PAPER

Parameter cascades and profiling in functional data analysis

Jiguo Cao · James O. Ramsay

Published online: 28 March 2007

Abstract A data smoothing method is described where the roughness penalty depends on a parameter that must be estimated from the data. Three levels of parameters are involved in this situation: *Local* parameters are the coefficients of the basis function expansion defining the smooth, *global* parameters define low-dimensional trend and the roughness penalty, and a *complexity* parameter controls the amount of roughness in the smooth. By defining local parameters as regularized functions of global parameters, and global parameters in turn as functions of complexity parameter, we define a parameter cascade, and show that the accompanying multi-criterion optimization problem leads to good estimates of all levels of parameters and their precisions. The approach is illustrated with real and simulated data, and this application is a prototype for a wide range of problems involving nuisance or local parameters.

Keywords Generalized cross-validation · Generalized profiled estimation · Nuisance parameters · Structural parameters · Markov Chain Monte Carlo

1 Introduction: Local, global and complexity parameters

Here is a line of enquiry that leads to an interesting question. In Green and Silverman (1994), Ramsay and Silverman (2002, 2005) and most situations requiring the flexible estimation of a functional parameter, we use basis function expansions of the form

J. Cao
Department of Mathematics and Statistics, McGill University,
805 Sherbrooke West, Montreal QC, Canada H3A 2K6

J. O. Ramsay (✉)
Department of Psychology, McGill University, 1205 Dr. Penfield Ave.,
Montreal QC, Canada H3A 1B1
e-mail: ramsay@psych.mcgill.ca

$$x(t) = \sum_{k}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t). \tag{1}$$

Our own preference has been to use at least as many basis functions K as data values $(t_i, y_i), i = 1, \ldots, n$, so that, in principle, any amount of variation in the data can be captured by the expansion. For example, $\phi_k(t)$ is often a B-spline basis function defined by placing knots at each observed argument value t_j, $j = 1, \ldots, n$.

Of course, using so many basis functions would lead to over-fitting of the data, unless we rely on roughness penalties of the form

$$P(x) = \lambda \int [Lx(t)]^2 \, dt$$

where L in our work has been a linear differential operator of order m of the form

$$Lx(t) = \sum_{j=0}^{m-1} \beta_j(t) D^j x(t) + D^m x(t).$$

That is, we minimize

$$\begin{aligned} J(\mathbf{c}|\boldsymbol{\beta}, \lambda) &= \sum_{j}^{n} [y_j - x(t_j)]^2 + \lambda \int [Lx(t)]^2 \, dt \\ &= \sum_{j}^{n} [y_j - \mathbf{c}'\boldsymbol{\phi}(t_j)]^2 + \mathbf{c}'\mathbf{R}(\lambda)\mathbf{c}. \end{aligned} \tag{2}$$

where

$$\mathbf{R}(\lambda) = \lambda \int L\boldsymbol{\phi}(t) L\boldsymbol{\phi}(t)' \, dt. \tag{3}$$

While most analyses employ simple differential operators of the form $Lx = D^m x$, we have found many good reasons to work with more sophisticated operators, and the final chapters of Ramsay and Silverman (2005) are devoted to treating the operator L as an object to be estimated from the data.

Figure 1 shows the incidence melanoma in the state of Connecticut over the 37 years beginning in 1937. We see sinusoidal and linear trends superimposed, and if we define

$$L_4 x(t) = e^{2\beta} D^2 x(t) + D^4 x(t), \tag{4}$$

then any trend of the form

$$x_0(t) = w_1 + w_2 t + w_3 \sin(e^{\beta} t) + w_4 \cos(e^{\beta} t) \tag{5}$$

in $x(t)$ is set to 0 as $\lambda \to \infty$, and the fit to the data tends to $x_0(t)$, which is a solution of the differential equation $L_4 x(t) = 0$. We need to estimate from the data the constant

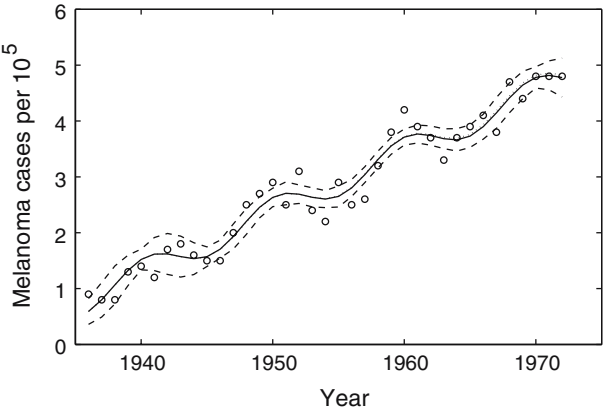

Fig. 1 The *circles* are the number of Age-adjusted incidences of melanoma per 10^5 from 1936 to 1972. The *dotted curve* is the solution of the linear differential equation $D^4x = -e^{2\beta}D^2x$ when $\beta = -0.445$. The *solid curve*, almost covering the dotted curve, is the *fitted curve* estimated by penalized smoothing with the penalty term defined by (4), when $\beta = -0.445$ and the smoothing parameter $\lambda = 2.4 * 10^4$

$\beta = \ln(2\pi/P)$, where P is the period of the sunspot cycle that causes the harmonic behavior in melanoma incidence. Figure 1 also shows the fits to the data defined by minimizing (2) with λ estimated by the profiling process described in Section 2.

Alternatively, we might also consider the semi-parametric model

$$J(\mathbf{c}|\boldsymbol{\beta}, \lambda) = \sum_{j}^{n}[y_j - \alpha_0 - \alpha_1 t - x(t_j)]^2 + \lambda \int [L_2 x(t)]^2 \, dt, \tag{6}$$

where the order two differential operator in the penalty term is

$$L_2 x(t) = e^{2\gamma} x(t) + D^2 x(t). \tag{7}$$

As $\lambda \to \infty$, the fit to the data tends to $\alpha_0 + \alpha_1 t + x_1(t)$, where

$$x_1(t) = w_3 \sin(e^{\gamma} t) + w_4 \cos(e^{\gamma} t) \tag{8}$$

is a solution of the differential equation $L_2 x(t) = 0$. Both models have the capacity to model tilted sinusoidal trend plus extra variation, but this models requires less differentiability in the solution $x(t)$ and introduces two additional parameters.

This approach of using "designer" penalty terms might seem a little more sophisticated than many functional data analyses require. Why did we adopt it? First of all, we and others have found that the "saturated basis plus roughness penalty" strategy produces better estimates of functions and their derivatives than commonly used alternatives, such as kernel and local polynomial smoothing, and certainly far better estimates than simple least squares fitting of a small and fixed number of basis functions. This is not to say that a simpler approach won't often do the job; but when we needed to push the data as far as possible, we have not found anything better.

Secondly, we discovered that choosing a roughness penalty that is estimated from the data opened up powerful new techniques and new opportunities for statistical approaches to data analysis. By smart choices of roughness penalties, we found that we could get better estimates of functional parameters and their derivatives. Moreover, since a linear differential operator is just another format for an linear ordinary differential equation, we saw this as a way to model data with differential equations. Our recent work Ramsay et al. (2007) has extended the material in our books to systems of nonlinear differential equations and we are now working on methods for partial differential equations.

Now here's the question. What sort of parameters are these coefficients c_k that define the basis function expansion (1)?

Their number K does not seem to be fixed since the more data that we collect, the larger K will be. Consequently, it seems inappropriate to use classical estimation theory that relies on the sample size becoming arbitrarily larger than the number of parameters. Moreover, the role of these coefficients depends on other incidental or design properties of the data, such as the spacing of the t_j's, the range of t-values over which the data are observed, and so forth. This can make comparing results obtained from different samples, different investigators and different laboratories difficult.

These coefficients are *nuisance* or *incidental parameters*, as opposed to *structural* parameters, terms that were proposed in Neyman and Scott (1948). In fact, the coefficients c_k are *data-proxies* in the sense that, if a saturated B-spline basis is used, then each coefficient represents the typical size of the observations y_i over the bounded interval over which the kth basis function is nonnegative. The situation with respect to a Fourier basis is similar, except that the local data property is expressed in the frequency domain. It seems descriptive to also use the term *local parameters* for parameters of this nature.

Structural parameters, on the other hand, are typically fixed in number no matter how much data are collected and, moreover, changing any of their values has an impact on almost all of the fit $x(t)$. Parameter β in (2) and parameters $(\alpha_0, \alpha_1, \gamma)$ in (6) are in this class, and we may prefer the term *global parameters* for these.

To bring this discussion into sharper focus, consider using an order 6 B-spline basis system for model (2) defined by placing knots at each of the sampling points t_j, which defines $n + 4 = 41$ basis functions in vector $\boldsymbol{\phi}(t)$. Define the n by $n + 4$ matrix $\boldsymbol{\Phi}$ as containing the values of the basis functions at the times t_j. Then the coefficient vector $\hat{\mathbf{c}}$ minimizing the criterion $\mathbf{J}(\mathbf{c}|\beta, \lambda)$ is

$$\hat{\mathbf{c}}(\beta, \lambda) = [\boldsymbol{\Phi}'\boldsymbol{\Phi} + \mathbf{R}(\beta, \lambda)]^{-1}\boldsymbol{\Phi}'\mathbf{y} \tag{9}$$

where

$$\mathbf{R}(\beta, \lambda) = \lambda \int L_4\boldsymbol{\phi}(t) L_4\boldsymbol{\phi}(t)' \, dt. \tag{10}$$

This conditional minimization defines the local or nuisance parameters as a vector-valued function of the global or structural parameters, and consequently reduces the dimensionality of the parameter space down to the number of structural parameters, in this case one. In this process, the local parameters lose their status as independent parameters. This process of defining a subset of the parameters as functions of other

parameters is known as *profiling*. The profile method has been investigated by many authors, see Severini and Wong (1992), Severini and Staniswalis (1994), Murphy and van der Vaart (2000) and their references. Recently, Keilegom and Carroll (2006) studied the asymptotic distribution of the profile estimate. However, Hung and Wong (1999) gave one example that the profile estimate "can be misleading in both precision (degree of freedom) and location (bias) especially in small-sample problems" if the criteria were same in all three levels, since the profile likelihood is not a true likelihood function.

Similarly, we can use an order 4 B-spline basis system with knots at sampling points for model (6) and drop the initial and final basis functions, so that there are 37 local parameters and 3 global parameters.

We now estimate β in (2) by optimizing the *un-penalized criterion*

$$\begin{aligned} H(\beta|\lambda) &= \sum_{j}^{n}[y_j - \hat{x}(t_j, \beta|\lambda)]^2 = \sum_{j}^{n}[y_j - \hat{\mathbf{c}}(\beta, \lambda)'\boldsymbol{\phi}(t_j)]^2 \\ &= \mathbf{y}'[I - \mathbf{A}(\beta|\lambda)]'[I - \mathbf{A}(\beta|\lambda)]\mathbf{y} \end{aligned} \tag{11}$$

where the $n \times n$ smoothing matrix $\mathbf{A}(\beta|\lambda) = \boldsymbol{\Phi}[\boldsymbol{\Phi}'\boldsymbol{\Phi} + \mathbf{R}(\beta|\lambda)]^{-1}\boldsymbol{\Phi}'$. We drop the roughness penalty here because the fitting function is already smoothed by virtue of minimizing $J(\mathbf{c}|\beta, \lambda)$, and so does not need regularizing twice. This procedure departs from the more usual *joint estimation strategy* in which the penalized least squares criterion $J(\mathbf{c}, \beta|\lambda)$ is minimized with respect to both $\mathbf{c}$ and β. Our earlier experience with this approach, reported in Heckman and Ramsay (2000), gave unsatisfactory estimates of both $\mathbf{c}$ and β. By minimizing the un-penalized criterion $H(\beta|\lambda)$, we obtain the estimate $\hat{\beta}$ for any value of λ, so $\hat{\beta}$ is an implicit function of λ.

There is, finally, the question of the status of λ. While it would appear to be a global or structural parameter, in fact it is not a part of the model in the sense that we attach no interpretive significance to its value. Instead, it's role is to control the overall complexity of the fit $x(t)$, and we might refer to it as a *complexity parameter*.

Smoothing parameter λ is often selected by minimizing the generalized cross-validated criterion, which can be expressed as

$$F(\hat{\beta}(\lambda), \lambda) = \mathtt{GCV}(\hat{\beta}(\lambda), \lambda) = n\frac{\mathtt{SSE}(\hat{\beta}(\lambda), \lambda)}{[\mathtt{dfe}(\hat{\beta}(\lambda), \lambda)]^2} \tag{12}$$

where degrees of freedom measure $dfe(\lambda)$ is

$$\mathtt{dfe}(\lambda) = n - \mathrm{trace}[\mathbf{A}(\lambda)]$$

and where the notation $F(\hat{\beta}(\lambda), \lambda|\mathbf{y})$ reminds us that F should be considered as a function of λ, since $\hat{\beta}$ is an implicit function of λ.

The criterion GCV is an error sum of squares discounted for the degrees of freedom invested in the model, and allows us to choose the λ value by optimizing a fit measure that is relatively insensitive to complexity. Efron (2004) and Gu (2002)

review the large literature on discounted fit measures and propose some interesting new approaches.

2 Parameter cascades and profiled parameter estimation

We have a distinct hierarchy in these three classes of parameters that one might refer to as a *parameter cascade*. At the bottom are the large number of local parameters in $\mathbf{c}$, and these are defined here as *functions* $\hat{\mathbf{c}}(\beta, \lambda)$ of the global and complexity parameters through the estimate (9) and its analogue for model (6). Even if an explicit solution for $\hat{\mathbf{c}}$ were not possible, we can still define this function *implicitly* by optimizing a penalized or regularized fitting function like $J(\mathbf{c}|\beta, \lambda)$ *each time* we change the value of a global parameter. When the functional relationship is implicit in this way, and the regularity conditions hold that are required to ensure that the optimization problem has a unique solution, the Implicit Function Theorem permits the explicit gradient and Hessian calculations that are essential for fast optimization.

The distinctive aspect of the parameter cascade as described in this paper is that a different criterion is optimized for each of the three levels of the cascade. That is, we solve a three-criterion optimization problem, where the criteria are $J(\mathbf{c}|\beta, \lambda)$, $H(\beta|\lambda)$ and $F(\lambda)$ for the local, global and complexity parameters, respectively.

The process that we have outlined is a generalization of the technique of profiling that is often used in nonlinear least squares and other fitting situations. When used as a computational trick, profiling defines a subset of parameters that are easily estimated as a function of the remaining parameters for which no simple closed form estimate is possible. This can often speed up computation, but does not affect the final estimated parameter values since the same fitting criterion is used at all stages of the cascade. That is, the essential difference between joint estimation and our approach is in the use of different fitting criteria at each level. In fact, the *joint estimation* of λ, β and $\mathbf{c}$ through optimizing $J(\mathbf{c}|\beta, \lambda)$ would lead to a reduction of the criterion to zero and the fitting of the data by an interpolating spline, and the strategy of choosing λ to minimize $F(\lambda)$ used in many smoothing spline applications works precisely because $F(\lambda)$ and $J(\mathbf{c}|\beta, \lambda)$ are not the same criteria, the former assessing mean squared error in the parameter estimates, and the latter assessing roughness penalized data fit.

We now abstract this process so as to be able to apply it more widely. Let $\mathbf{c}$ be a vector of local or nuisance parameters, let $\boldsymbol{\beta}$ be a vector of global or structural parameters, and let $\boldsymbol{\theta}$ be a vector of complexity parameters.

Let $\mathbf{J}(\mathbf{c}|\beta, \theta)$ be an *inner* criterion. It will typically be based on an error sum of squares, log likelihood, posterior density or any suitable measure of the quality of the fit to the data plus a regularization or smoothing term that defines smoothness in terms of $\mathbf{c}$, and it will also depend on complexity parameter $\boldsymbol{\theta}$ and possibly also on one or more of the global parameters in $\boldsymbol{\beta}$. The nuisance parameter vector $\mathbf{c}$ is removed from the parameter space by defining the inner optimization conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Let $H(\mathbf{c}(\boldsymbol{\beta}), \boldsymbol{\beta}|\boldsymbol{\theta})$ be a *middle* criterion that defines fit to the data conditional on $\boldsymbol{\theta}$, and this may also be regularized, but the regularization term will define smoothness as a function of $\boldsymbol{\beta}$ rather than as a function of $\mathbf{c}$.

Let $F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ be an *outer* criterion, that can be based on generalized cross validation, or any suitable measure of the model complexity. The profiling process requires that the middle criterion H is optimized with respect to $\boldsymbol{\beta}$ each time $\boldsymbol{\theta}$ is changed, so that the estimating function $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is defined implicitly by this profiling strategy, and, if we are lucky, explicitly as well, but this is not essential.

As a consequence of these conditional optimizations, the nuisance parameter vector $\mathbf{c}$ is removed from the parameter space as an independent parameter by defining it through the inner optimization as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. In the same way, the structural parameter vector $\boldsymbol{\beta}$ is removed from the parameter space through the middle optimization as a function of $\boldsymbol{\theta}$. Our final parameter estimates become the functional cascade $\hat{\mathbf{c}}[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})]$, $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\theta}}$ defined by the optimizations with respect to criteria $\mathbf{J}$, $\mathbf{H}$ and $\mathbf{F}$, respectively.

The optimization of $F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ becomes faster and more stable if we have the gradient

$$\frac{dF(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{\partial F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{\partial F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \hat{\boldsymbol{\beta}}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}}, \tag{13}$$

where $dF(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})/d\boldsymbol{\theta}$ is the total derivative of F with respect to $\boldsymbol{\theta}$. Notice that the formula of $dF(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})/d\boldsymbol{\theta}$ involves the term $\partial \hat{\boldsymbol{\beta}}/\partial \boldsymbol{\theta}$. If the middle optimization leads to an explicit solution for $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, the gradient is readily available. But if not, the Implicit Function Theorem can be applied to find $\partial \hat{\boldsymbol{\beta}}/\partial \boldsymbol{\theta}$. Since the optimal local parameter vector $\hat{\boldsymbol{\beta}}$ satisfies $\partial H(\boldsymbol{\beta}|\boldsymbol{\theta})/\partial \boldsymbol{\beta} = 0$, and $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$, we can take the $\boldsymbol{\theta}$-derivative on $\partial H(\boldsymbol{\beta}|\boldsymbol{\theta})/\partial \boldsymbol{\beta}|_{\hat{\boldsymbol{\beta}}} = 0$ as follows:

$$\frac{d}{d\boldsymbol{\theta}} \left(\frac{\partial H(\boldsymbol{\beta}|\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \bigg|_{\hat{\boldsymbol{\beta}}} \right) = \frac{\partial^2 H(\boldsymbol{\beta}|\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} \bigg|_{\hat{\boldsymbol{\beta}}} + \frac{\partial^2 H(\boldsymbol{\beta}|\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^2} \bigg|_{\hat{\boldsymbol{\beta}}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} = 0\,, \tag{14}$$

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

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} = -\left[\frac{\partial^2 H(\boldsymbol{\beta}|\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^2} \bigg|_{\hat{\boldsymbol{\beta}}} \right]^{-1} \left[\frac{\partial^2 H(\boldsymbol{\beta}|\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} \bigg|_{\hat{\boldsymbol{\beta}}} \right]. \tag{15}$$

Further results when F is the GCV criterion (12) are provided in the Appendix.

3 Interval estimation for nuisance, structural and complexity parameters

In this section, we derive the variances for nuisance, global and complexity parameters by modifying the Delta method. By treating global parameters as functions of complexity parameters, the variances of global parameters also include the variation inherited

from the complexity parameters. Let $\boldsymbol{\Sigma}$ denote the variance–covariance matrix for $\mathbf{y}$, which can be estimated in the smoothing context by:

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathtt{SSE}(\hat{\boldsymbol{\theta}})}{\mathtt{dfe}(\hat{\boldsymbol{\theta}})} \cdot \mathbf{I}. \tag{16}$$

The estimated complexity parameter vector $\hat{\boldsymbol{\theta}}$ satisfies $\partial F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \mathbf{y})/\partial \boldsymbol{\theta} = 0$. By taking the $\mathbf{y}$-derivative of $\partial F(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}, \mathbf{y})/\partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} = 0$, we obtain:

$$\frac{d}{d\mathbf{y}}\left(\frac{dF}{d\boldsymbol{\theta}}\bigg|_{\hat{\boldsymbol{\theta}}, \mathbf{y}}\right) = \frac{d^2F}{d\boldsymbol{\theta} d\mathbf{y}}\bigg|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} + \frac{d^2F}{d\boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}}, \mathbf{y}} \frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = 0, \tag{17}$$

where

$$\frac{d^2F}{d\boldsymbol{\theta}^2} = \frac{\partial^2 F}{\partial \boldsymbol{\theta}^2} + 2\frac{\partial^2 F}{\partial \hat{\boldsymbol{\beta}} \partial \boldsymbol{\theta}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}}\right)' \frac{\partial^2 F}{\partial \hat{\boldsymbol{\beta}}^2} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial F}{\partial \hat{\boldsymbol{\beta}}}\right)^2 \frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}^2}, \tag{18}$$

and

$$\frac{d^2F}{d\boldsymbol{\theta} d\mathbf{y}} = \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\boldsymbol{\beta}} \partial \mathbf{y}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \hat{\boldsymbol{\beta}}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\boldsymbol{\beta}}^2} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \mathbf{y}} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial F}{\partial \hat{\boldsymbol{\beta}}}\right)^2 \frac{\partial^2 \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\theta} \partial \mathbf{y}}. \tag{19}$$

Equations (17) holds since $\partial F/\partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}, \mathbf{y}}$ is a function of $\mathbf{y}$ that is identically 0.

Solving Eq. (17), we get the first derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = -\left[\frac{d^2F}{d\boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}}, \mathbf{y}}\right]^{-1} \left[\frac{d^2F}{d\boldsymbol{\theta} d\mathbf{y}}\bigg|_{\hat{\boldsymbol{\theta}}, \mathbf{y}}\right]. \tag{20}$$

Letting $\boldsymbol{\mu} = \mathtt{E}(\mathbf{y})$ and using its first order Taylor expansion, we have that

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) \approx \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) + \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu}). \tag{21}$$

Consequently, the variance of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ can be estimated by

$$\mathtt{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right] \boldsymbol{\Sigma} \left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right]' \approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right] \boldsymbol{\Sigma} \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]'. \tag{22}$$

Approximation (22) makes sense since

$$\mathtt{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right) \approx \mathtt{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right), \tag{23}$$

when $d^2\hat{\boldsymbol{\theta}}/d^2\boldsymbol{\mu}$ are bounded by a fixed number, which can be derived by taking expectation on both sides of the first order Taylor expansion for $d\hat{\boldsymbol{\theta}}/d\mathbf{y}$:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} \approx \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}} + \frac{d^2\hat{\boldsymbol{\theta}}}{d^2\boldsymbol{\mu}}(\mathbf{y}-\boldsymbol{\mu}). \tag{24}$$

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

$$\mathtt{Var}[\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})] \approx \left[\frac{d\hat{\boldsymbol{\beta}}}{d\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{d\hat{\boldsymbol{\beta}}}{d\mathbf{y}}\right]', \tag{25}$$

where

$$\frac{d\hat{\boldsymbol{\beta}}}{d\mathbf{y}} = \frac{\partial\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\theta}}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{y}}. \tag{26}$$

The sampling variance of $\hat{\mathbf{c}}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y}), \hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})$ is estimated by

$$\mathtt{Var}[\hat{\mathbf{c}}(\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y}), \hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})] \approx \left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right]', \tag{27}$$

where

$$\frac{d\hat{\mathbf{c}}}{d\mathbf{y}} = \frac{\partial\hat{\mathbf{c}}}{\partial\hat{\boldsymbol{\theta}}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\hat{\boldsymbol{\beta}}}\frac{d\hat{\boldsymbol{\beta}}}{d\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}. \tag{28}$$

If we do not consider the functional relationship between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, the sampling variance of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}(\mathbf{y}))$ will then be underestimated by replacing the full derivative of $\hat{\boldsymbol{\beta}}$ with respect to $\mathbf{y}$ by the partial derivative of $\hat{\boldsymbol{\beta}}$ with respect to $\mathbf{y}$:

$$\mathtt{Var}[\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}}, \mathbf{y}] \approx \left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\mathbf{y}}\right]'. \tag{29}$$

We call $\mathtt{Var}[\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}}, \mathbf{y}]$ *the conditional sampling variance* for $\hat{\boldsymbol{\beta}}$, because it ignores the uncertainty resulting from the estimate $\hat{\boldsymbol{\theta}}$. Similarly, the conditional sampling variance for $\hat{\mathbf{c}}$ is

$$\mathtt{Var}[\hat{\mathbf{c}}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \mathbf{y}] \approx \left[\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\right]', \tag{30}$$

4 Profiled estimation applied to the melanoma data

4.1 Results for the fourth order smoothing criterion (2)

In order to make sure to get a positive smoothing parameter λ, we estimate $\theta = \ln(\lambda)$ instead. For the melanoma data shown in Fig. 1, we estimate $\hat{\theta} = 10.1$ with $\mathtt{STD}(\hat{\theta}) = 0.81$, and $\hat{\beta} = -0.445$ with $\mathtt{STD}(\hat{\beta}) = 0.062$. The dotted curve in Fig. 1 is the solution of the linear differential equation (ODE) $D^4x = -\exp(2\beta)D^2x$ when $\beta = -0.445$.

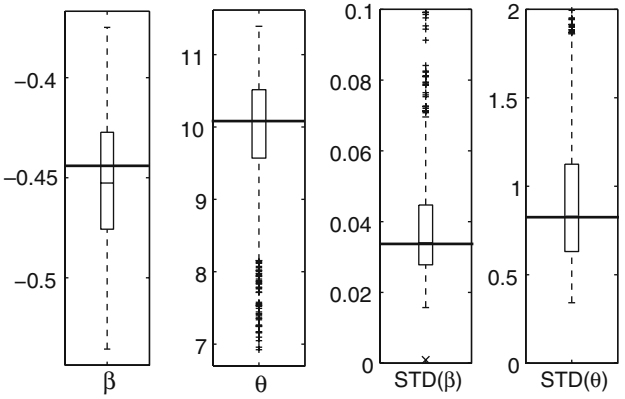

Fig. 2 The *boxplots* for the estimated β, $\theta = \log(\lambda)$ and their standard deviations. The *long horizontal bars* in the first two are the true values that generate the simulated data sets, and the *horizontal bars* in the third and fourth graphs are the sample standard deviations for β and θ. Notice that the medians in the last three *boxplots* are almost covered by the *long horizontal bars*. The cross in the third graph is the median of the conditional standard deviation for β

Figure 1 also displays the fitted curve with its pointwise 95% confidence interval. We can see the ODE solution is very close to the fitted curve, which indicates that this ODE describes well the dynamical behavior of melanoma incidence. This can also be verified by the large value of the estimate $\hat{\lambda} = e^{10.1}$ for the smoothing parameter. The estimated variance of the observations is $\hat{\sigma}^2 = 0.070$.

To verify the accuracy of our estimates of parameters and their standard errors, we generated 1000 simulated data sets by adding white noise with variance $\sigma^2 = 0.070$ to the differential equation solution with $\beta = -0.445$. Figure 2 contains boxplots for the estimated β and θ as well as their respective standard deviations. The biases for $\hat{\beta}$'s and $\hat{\theta}$ are 1.6% and 5% of their true values, respectively. The long lower tail in the boxplot for $\hat{\theta}$ indicates that GCV sometimes seriously underestimates the smoothing parameters, which is a known feature of the criterion (Gu 2002). The medians of the unconditional standard deviation estimates for $\hat{\beta}$ and $\hat{\theta}$ are almost the same as their corresponding sample standard deviations. In comparison, the median of the conditional standard deviation estimates for $\hat{\beta}$ is far below the sample standard deviation, indicating that the precision of $\hat{\beta}$ will be seriously underestimated if the dependency of $\hat{\theta}$ on the data is not taken into account.

Figure 3 shows that the median of the unconditionally estimated pointwise standard deviation of $x(t)$ from these simulated data sets is close to the sample standard deviation. Both of these standard deviations are much larger than the median of the conditional estimates, which seriously underestimate the variability of $\hat{x}(t)$ by ignoring the data dependency of $\hat{\beta}$ and $\hat{\theta}$.

4.2 Profiled estimation for the semi-parametric model (6)

All of the calculations are similar as for the fourth order smoothing criterion (2), except that the coefficient vector $\hat{\mathbf{c}}$ minimizing the criterion (6) is

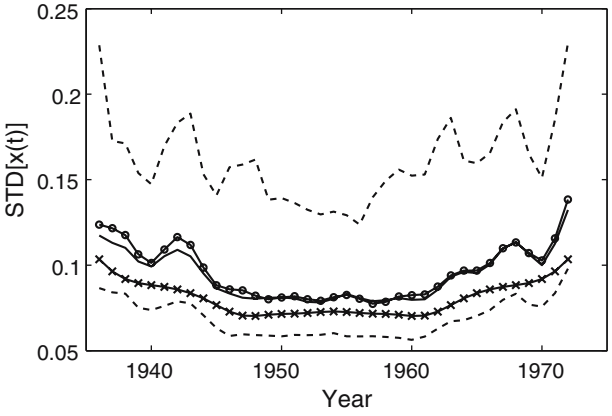

Fig. 3 The *solid line* is the median of the unconditionally estimated standard deviation of $x(t)$ from these 1000 simulated data sets. The *dashed lines* are the the 2.5%, 97.5% quantiles of the estimates. The *solid line* marked with *circles* is the sample standard deviation of $x(t)$, and the *solid line* marked with *crosses* is the median of the conditional standard deviation estimates for $x(t)$

$$\hat{\mathbf{c}}(\boldsymbol{\beta}, \lambda) = [\boldsymbol{\Phi}'\boldsymbol{\Phi} + \mathbf{R}^*(\gamma, \lambda)]^{-1}\boldsymbol{\Phi}'(\mathbf{y} - \mathbf{M}\boldsymbol{\alpha}), \tag{31}$$

where $\mathbf{R}^*(\gamma, \lambda) = \lambda \int L_2\boldsymbol{\phi}(t) L_2\boldsymbol{\phi}(t)' dt$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$, and $\mathbf{M}$ is the $n \times 2$ design matrix with the ith row to be $(1, i)$.

From the real data shown in Fig. 4, we obtain the statistical inferences of the parameters in the semi-parametric model (6), which are shown in Table 1. The estimated data variance is 0.067, which is calculated by $\hat{\sigma}^2 = \texttt{SSE}/(\texttt{dfe} - 2)$. Here the degrees of freedom is set of $\texttt{dfe} - 2$, since we have two linear coefficients α_0 and α_1. Figure 4 displays the predicted curve

$$\hat{y}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\mathbf{c}}' * \boldsymbol{\phi}(t) \tag{32}$$

along with its estimated 95% confidence interval.

To verify our estimates, we generate 1000 simulated data sets by adding white noise with variance $\sigma^2 = 0.067$ to the predicted curve (32). Figure 5 illustrates the boxplots for the estimates of α_0, α_1, γ and $\theta = \log(\lambda)$. The estimates of α_0, α_1 are almost unbiased, and the biases for the estimates of γ and θ are only 0.5% and 2% of their true values.

Figure 6 shows the boxplots for the standard deviation estimates of these four parameters, the median of which are close to their sample standard deviations. As we would expect, the median of the conditional standard deviation estimates for β is far lower than the sample variance, which means the standard deviation for β is underestimated by the conditional estimates.

Figure 7 displays the 2.5%, 50%, 97.5% quantiles of the estimated standard deviation for $x(t)$. The median of the estimated standard deviation for $x(t)$ is close to the sample standard deviation, but the median of the conditional estimate is far lower than both of them.

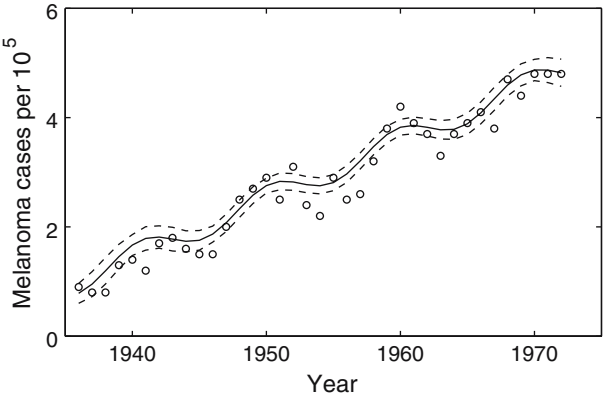

Fig. 4 The *circles* are the number of Age-adjusted incidences of melanoma per 10^5 from 1936 to 1972. The *solid curve* is the predicted curve $\hat{y}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\mathbf{c}}(\gamma)'\boldsymbol{\phi}(t)$, with the 95% confidence interval given as *dashed lines*

Table 1 The statistical inferences of the parameters in the semi-parametric model (6)

Parameters	Estimates	STD's
α_0	0.92	8.8×10^{-2}
α_1	0.11	4.0×10^{-3}
β	−0.43	3.1×10^{-2}
θ	7.5	0.56

5 Discussion and conclusions

The two models for Melanoma data involve two differential equations with analytical solutions, so it is certainly possible to work on their solutions (5) straightaway. But usually the forms of differential equations are much simpler than their analytical solutions. Moreover, a large number of models in engineering, ecology and many other areas are given directly in the form of differential equations, and most of them don't have analytical solutions. In fact, estimating differential equations is a crucial problem, which is also called *inverse problem in engineering*.

Wherever parameters are tied to local characteristics of the data, and consequently are large in number relative to other parameters requiring estimation, it is natural to consider some sort of regularization of their estimates. This leads in turn to an inner and an outer optimization, where in the inner loop conditional regularized local parameter estimates are computed, and in the outer loop the global parameters are estimated. That is, the nuisance parameters are effectively functions of the global parameters, whether implicitly or explicitly. The coefficients of basis function expansions of functional parameters are natural candidates for this treatment.

The regularization process can itself depend on unknown parameters, as we illustrated with the melanoma data. In fact, regularization is often defined within a Bayesian framework, where the roughness penalty is the log of a prior density. In this case, the prior density is often a function of one or more parameters, such as prior variances and

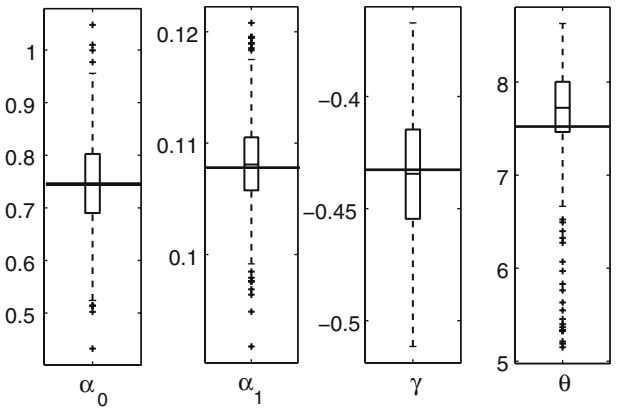

Fig. 5 The *boxplots* for the estimated α_0, α_1, γ and $\theta = \log(\lambda)$ from these 1000 simulated data sets. The *horizontal bars* are the true values of parameters that generate the simulated data sets. Notice that the median in the first boxplot is covered by the *horizontal bars*

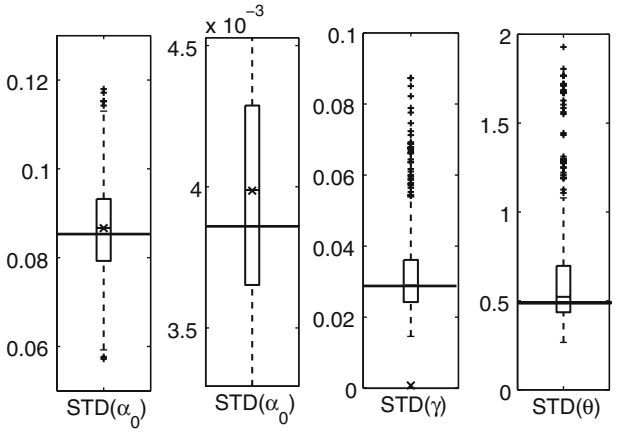

Fig. 6 The *boxplots* for the unconditionally estimated $\mathtt{STD}(\alpha_0)$, $\mathtt{STD}(\alpha_1)$, $\mathtt{STD}(\gamma)$ and $\mathtt{STD}(\theta)$ from 1000 simulated data sets. The *long horizontal bars* are their sample standard deviations, and the *cross* in the first three graphs is the median of conditional standard deviation estimates. Notice that the median in the third boxplot is covered by the *long horizontal bars*

covariances, and it is such parameters that we refer to here as complexity parameters. These can be fixed, as they often are, or they can be estimated from the data. In either case, we view the use of such *informative* regularizers as a powerful extension of the model building process. We have illustrated in our analysis of the melanoma data that the multi-criterion optimization process implied by a parameter cascade can lead to accurate estimates of both the parameters and of the precision of their estimates.

In recent years, statisticians have favored a Bayesian approach using Markov Chain Monte Carlo to remove local parameters by integrating over their values with respect to some prior measure. This idea is closely related to what we propose here, with the primary difference being that we use an optimizing estimate, which might be called

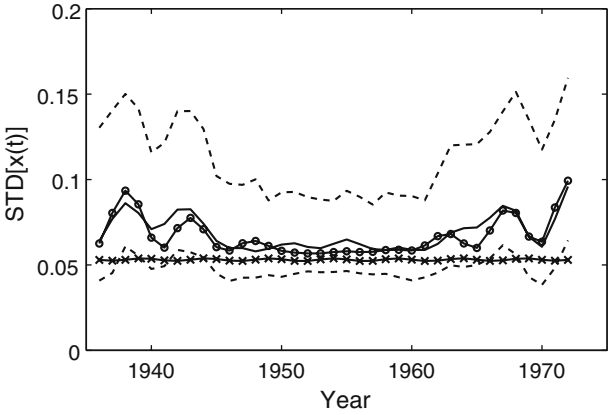

Fig. 7 The *solid line* is the median of the unconditionally estimated standard deviation of $x(t)$ from 1000 simulated data sets. The *dashed lines* are the the 2.5%, 97.5% quantiles of the estimates. The *solid line* marked with *circles* is the sample standard deviation of $x(t)$, and the *solid line* marked with *crosses* is the median of the conditional standard deviation estimates for $x(t)$

from a Bayesian perspective a *modal* estimate, rather than integration over potential values of nuisance parameters. However, this difference between the two approaches is less important than it may appear, since the mean value theorem states than an integral over $\mathbf{c}$ is equal to the range of integration times the integrand evaluated at some point $\hat{\mathbf{c}}$, and consequently this point may be thought of as implicitly defining a function of global parameter estimates derived from marginal likelihoods or posterior densities.

Although we do not want to diminish the importance and appeal of Bayesian methods in any way, nevertheless, the marginalization approach to nuisance parameters does have a number of serious drawbacks, and particularly with respect to deploying the technology to applied settings. The computer code required for MCMC methods must often be written in low level languages to achieve the required efficiency, and therefore represents a major programming investment for both the computational code and the user interface that envelopes it. Users are confronted with issues of the appropriate number of "burn-in" cycles, choices of jump distributions and other tuning considerations which they may be poorly equipped to understand, and that may be specific to the data being analyzed. The heavy computational burden of MCMC can also work against use in applications. Finally, marginalization as a concept is itself sometimes difficult to communicate to statistically naive users.

On the other hand, the extended profiling process works naturally with regularization approaches to control complexity, and casts the computational problem in a structural parameter space of often greatly reduced dimensionality. We have found the approach to be especially useful in the context of estimating differential equation models in Ramsay et al. (2007) where each output variable requires its own high dimensional basis expansion, but where the differential equation itself is defined by an often small number of structural parameters. We have found that estimates of these parameters and of their precisions can be obtained in a very small fraction of the computation time required by an MCMC-based algorithm used within a Bayesian framework.

The convenience of the multi-criterion optimization and parameter cascade concepts and the good statistical properties of parameter estimates obtained in this way suggest that this approach deserves consideration along with marginalization and other methods for dealing with local or nuisance parameters.

Appendix

The global parameter β and the complexity parameter θ in our model are estimated by the Newton–Raphson algorithm in the middle and outer optimization level, with the gradients and Hessian matrices given analytically. As a result, the computation is extremely fast, usually no more than 1 s. In the appendix, we supply the main mathematical formulas required in the outer optimization level. More mathematical details can be obtained by contacting the authors.

- The first derivative of $\mathtt{GCV}(\lambda(t))$ with respect to θ is

$$\frac{\partial \mathtt{GCV}(\lambda)}{\partial \theta} = n\left[\mathtt{dfe}\frac{\partial \mathtt{SSE}}{\partial \theta} - 2\mathtt{SSE}\frac{\partial \mathtt{dfe}}{\partial \theta}\right]\mathtt{dfe}^{-3} \tag{33}$$

  where

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

- The second derivative of $\mathtt{GCV}(\lambda(t))$ with respect to θ is

$$\begin{aligned}\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial \theta^2} = &\frac{n}{\mathtt{dfe}^2}\frac{\partial^2 \mathtt{SSE}}{\partial \theta^2} - \frac{2n\mathtt{SSE}}{\mathtt{dfe}^3}\frac{\partial^2 \mathtt{dfe}}{\partial \theta^2} + \frac{6n\mathtt{SSE}}{\mathtt{dfe}^4}\left[\frac{\partial \mathtt{dfe}}{\partial \theta}\right]^2\\ &-\frac{4n}{\mathtt{dfe}^3}\frac{\partial \mathtt{dfe}}{\partial \theta}\frac{\partial \mathtt{SSE}}{\partial \theta}\end{aligned} \tag{34}$$

  where

$$\begin{aligned}\frac{\partial^2 \mathtt{SSE}(\lambda)}{\partial \theta^2} &= \mathbf{y}'(E'+E)\mathbf{y}\\ \frac{\partial^2 \mathtt{dfe}(\lambda)}{\partial \theta^2} &= -\mathtt{Tr}\left(\frac{d^2\mathbf{A}}{d\theta^2}\right)\end{aligned}$$

  and

$$E = \left[\frac{d\mathbf{A}}{d\theta}\right]'\left[\frac{d\mathbf{A}}{d\theta}\right] - \left[\frac{d^2\mathbf{A}}{d\theta^2}\right]'[\mathbf{I}-\mathbf{A}].$$

- The second derivative of $\mathtt{GCV}(\lambda(t))$ with respect to θ and y_i is

$$\begin{aligned}\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial\theta\partial y_i} = n\Bigg[&\frac{\partial \mathtt{dfe}}{\partial y_i}\frac{\partial \mathtt{SSE}}{\partial\theta} + \mathtt{dfe}\frac{\partial^2 \mathtt{SSE}}{\partial\theta\partial y_i} \\ &-2\frac{\partial \mathtt{SSE}}{\partial y_i}\frac{\partial \mathtt{dfe}}{\partial\theta} - 2\mathtt{SSE}\frac{\partial^2 \mathtt{dfe}}{\partial\theta\partial y_i}\Bigg]\mathtt{dfe}^{-3} \\ &-3n\left[\left(\mathtt{dfe}\frac{\partial \mathtt{SSE}}{\partial\theta} - 2\mathtt{SSE}\frac{\partial \mathtt{dfe}}{\partial\theta}\right)\mathtt{dfe}^{-4}\frac{\partial \mathtt{dfe}}{\partial y_i}\right] \qquad (35)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial \mathtt{dfe}(\lambda)}{\partial y_i} &= -\mathtt{Tr}\left(\frac{d\mathbf{A}}{dy_i}\right) \\ \frac{\partial^2 \mathtt{dfe}(\lambda)}{\partial\theta\partial y_i} &= -\mathtt{Tr}\left(\frac{d^2\mathbf{A}}{d\theta dy_i}\right) \\ \frac{\partial \mathtt{SSE}(\lambda)}{\partial y_i} &= 2[(\mathbf{I}-\mathbf{A})'(\mathbf{I}-\mathbf{A})\mathbf{y}]_i - \mathbf{y}'\left[\left(\frac{d\mathbf{A}}{dy_i}\right)'(\mathbf{I}-\mathbf{A}) + (\mathbf{I}-\mathbf{A})'\left(\frac{d\mathbf{A}}{dy_i}\right)\right]\mathbf{y} \\ \frac{\partial^2 \mathtt{SSE}(\lambda)}{\partial\theta\partial y_i} &= -2\left[\left(\left(\frac{d\mathbf{A}}{d\theta}\right)'(\mathbf{I}-\mathbf{A}) + (\mathbf{I}-\mathbf{A})'\left(\frac{d\mathbf{A}}{d\theta}\right)\right)\mathbf{y}\right]_i \\ &\quad -\mathbf{y}'\left[\left(\frac{d^2\mathbf{A}}{d\theta dy_i}\right)'(\mathbf{I}-\mathbf{A}) - \left(\frac{d\mathbf{A}}{dy_i}\right)'\left(\frac{d\mathbf{A}}{d\theta}\right)\right. \\ &\quad \left.-\left(\frac{d\mathbf{A}}{d\theta}\right)'\left(\frac{d\mathbf{A}}{dy_i}\right) + (\mathbf{I}-\mathbf{A})'\left(\frac{d^2\mathbf{A}}{d\theta dy_i}\right)\right]\mathbf{y}.\end{aligned}$$

Here, the notation $[]_i$ means the $i-$th entry in the vector inside [].

References

Efron B (2004) The estimation of prediction error: covariance penalties and cross-validation. J Am Stat Assoc 99:619–642

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall, London

Gu C (2002) Smoothing spline ANOVA models. Springer, New York

Heckman N, Ramsay J (2000) Penalized regression with model based penalties. Can J Stat 28:241–258

Hung H, Wong W (1999) Averaging and profiling of likelihoods and the nuisance parameter problem. Technical report, Department of Statistics, Stanford University

Keilegom IV, Carroll RJ (2006) Backfitting versus profiling in general criterion functions. Statistica Sinica (submitted)

Murphy SA, van der Vaart AW (2000) On profile likelihood. J Am Stat Assoc 95:449–485

Neyman J, Scott EL (1948) Consistent estimates based on partially consistent observations. Econometrika 16:1–32

Ramsay JO, Silverman BW (2002) Functional data analysis, 1st edn. Springer, New York

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York

Ramsay JO, Hooker G, Campbell D, Cao J (2007) Estimating differential equations (with discussion). J R Stat Soc Ser B (in press)

Severini T, Staniswalis J (1994) Quasi-likelihood estimation in semiparametric models. J Am Stat Assoc 89:501–511
Severini T, Wong WH (1992) Profile likelihood and conditionally parametric models. Ann Stat 20:1768–1802