Stat Comput (2012) 22:857–865
DOI 10.1007/s11222-011-9252-1

Estimating generalized semiparametric additive models using parameter cascading

Jiguo Cao

Received: 30 March 2010 / Accepted: 21 January 2011 / Published online: 5 April 2011

Abstract Elimination of nuisance parameters is a central but difficult problem in statistical inference. We propose the parameter cascading method to estimate statistical models that involve nuisance parameters, structural parameters, and complexity parameters. The parameter cascading method has several unique aspects. First, we consider functional relationships between parameters, quantitatively described using analytical derivatives. These functional relationships can be explicit or implicit, and in the latter case the Implicit Function Theorem is applied to obtain the required derivatives. Second, we can express the gradients and Hessian matrices analytically, which is essential for fast and stable computation. Third, we develop the unconditional variance estimates for parameters, which include the uncertainty coming from other parameter estimates. The parameter cascading method is demonstrated by estimating generalized semiparametric additive models (GSAMs), where the response variable is allowed to be from any distribution. The practical necessity and empirical performance of the parameter cascading method are illustrated through a simulation study, and two applied example, one on finding the effect of air pollution on public health, and the other on the management of a retirement fund. The results demonstrate that the parameter cascading method is a good alternative to traditional methods.

Keywords Penalized smoothing · Nuisance parameters · Air pollution

J. Cao (✉)
Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada
e-mail: jiguo_cao@sfu.ca

1 Introduction

Many statistical models have some parameters not of direct interest, which are called *nuisance parameters*. The number of nuisance parameters is often not fixed, sometimes even increasing linearly with the sample size. Consequently, it seems inappropriate to use classical estimation theory which relies on the sample size becoming arbitrarily larger than the number of parameters. How to estimate statistical models with nuisance parameters is a central but difficult problem in statistical inference.

There are two popular likelihood approaches to address the nuisance parameter problem. Bayesian integrated likelihood methods obtain the marginal posterior distribution of the structural parameters by integrating the joint posterior distribution over the nuisance parameters (Berger et al. 1999). Profiling the likelihood is another standard approach to eliminate nuisance parameters. But since the profile likelihood is not a true likelihood, the parameter estimates can be biased in many examples (Neyman and Scott 1948; Cruddas et al. 1989). Among many methods to correct the profile likelihood, the modified profile likelihood (Barndorff-Nielsen 1983) and the conditional profile likelihood (Cox and Reid 1987) provide approximations to the likelihood functions for the parameters of interest.

We develop the parameter cascading method to estimate statistical models with nuisance parameters. Our approach proceeds using nested levels of optimization to fit the model. Parameters at each level of these optimizations depend on parameters estimates from other levels of optimization. Our method explicitly accounts for the interdependence of these parameters, which leads to reduced bias and more accurate variance estimates. We develop the unconditional variance estimates for parameters, which include the uncertainty coming from the estimates of other parameters.

We apply the parameter cascading method to estimate the generalized semiparametric additive model (GSAM). The GSAM is an increasingly popular version of the generalized additive model in which one or more of the covariates are modeled parametrically (Ruppert et al. 2003; Härdle et al. 2004). Many methods have been proposed to fit GSAMs. Perhaps the simplest is weighted least squares (Zeger and Diggle 1994; Lin and Carroll 2001; Lin and Ying 2001; Fan and Li 2004), which is appropriate when the errors follow a Gaussian distribution. Severini and Staniswalis (1994) estimated GSAMs using a quasi-likelihood function and developed asymptotic distributions for their estimators. Lin and Carroll (2006) considered a wide class of semiparametric problems and proposed profile and back-fitting estimation methods based on kernel smoothing and showed that profiling and back-fitting had identical limit distributions. The GSAM has also been described under a variety of alternative names, including semiparametric model (Zeger and Diggle 1994; He et al. 2002), partially linear model (Speckman 1988; Wolfgang et al. 2000) and partially splined model (Rice 1986). The primary reason for the popularity of GSAMs is that they allow an analyst to flexibly control for unknown confounding effects while still providing easily interpretable linear effects for the covariates of interest.

The main contributions of this paper are as follows. First, we consider the interdependence of the nuisance parameters, structural parameters and complexity parameters. When the dependence cannot be expressed explicitly, the Implicit Function Theorem is applied. Second, Lin and Carroll (2006) suggested computing the gradients by numerical differentiation, but pointed out that this would be difficult to implement numerically. We solve this problem by expressing the gradients as well as Hessian matrices analytically, which is crucial for fast and stable computation. Third, we develop the unconditional variance estimates for parameters, which include the uncertainty coming from other parameters. Fourth, our method allows the response variable from any distribution. Finally, a general package to estimate GSAMs with our method has been developed in the Matlab programming language, making use of functional data analysis software intended to compliment Ramsay and Silverman (2005). This package can be run automatically, without worrying about the basis system selection or the smoothing parameter choice. This package is also easy to extend to estimate other statistical models involving three groups of parameters by choosing appropriate optimization criteria.

Assuming observations y_j, $j = 1, \ldots, n$, have means $\mu_j = \mathrm{E}(y_j)$, we can write the GSAM as follows:

$$\eta_j = g(\mu_j) = \sum_{i=1}^{P} f_i(Z_{ij}) + \sum_{k=1}^{Q} \beta_k X_{kj}, \tag{1}$$

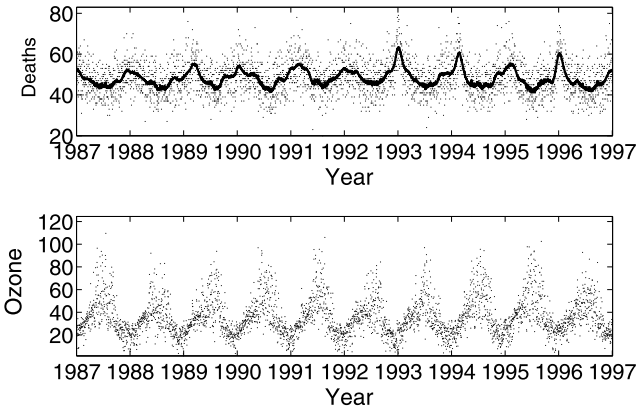

Fig. 1 The *top panel* displays the daily count of non-accidental deaths from 1987 to 1996 in Toronto, and the *bottom panel* shows the associate daily one-hour-maximum ozone

where $g(\cdot)$ is the link function. For instance, $g(\cdot)$ can be a log function for Poisson distributed observations or a logistic function for the binomial distributed data. Typically, X_k, $k = 1, \ldots, Q$, represents the variable or variables of interest, while Z_i, $i = 1, \ldots, P$, represents one or more confounding variables that the analyst would like to account for without making any restrictive parametric assumptions concerning the nature of their effects. X_{kj} and Z_{ij} are the j-th observation for X_k and Z_i. The functional parameters $f_i(Z_i)$ are fit using a nonparametric smoother. We consider the P functional parameters to be nuisance parameters and consider the linear coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_Q)$ to be the parameter of interest. If there is no linear part, then this model reduces to the familiar generalized additive model (Hastie and Tibshirani 1990; Wood 2006).

One important application of GSAMs involves time-series analysis of the acute effect of ambient air pollution on public health (Ramsay 2005); we will use this as our motivating and illustration. Figure 1 displays the daily counts of non-accidental deaths from 1987 to 1996 in Toronto, as well as the daily one-hour-maximum level of ozone. In our simple example, the objective is to determine whether the amount of daily ozone has any effect on mortality, after taking account of a seasonal trend, by fitting the following model:

$$\eta_j = \log(\mathrm{E}(y_j)) = f(D_j) + \beta P_j. \tag{2}$$

The integer variable D_j indexes the day and P_j represents the ozone level on the j-th day. The daily death count, represented by y_j, is assumed to be Poisson distributed. The functional parameter $f(D_j)$ represents a background time trend and is treated as a nuisance parameter. Only the structural parameter β, representing the effect of ozone on log-mortality, is of interest, and this effect is expected to be rather small in comparison with the background time trend.

This model has been heavily used in environmental epidemiology, usually with the addition of other confounding covariates such as temperature and day-of-the-week.

In 2002, the U.S. Environmental Protection Agency (EPA) was forced to delay publication of a review document on the health effects of airborne particulate matter when it was discovered that Model (2), on which much of the review document was based, had been giving biased estimates. An important source of bias is concurvity, or the fact that the background seasonal trend is confounded with the seasonal trend of the air pollutant. As a result of this confounding, the fitted pollution effect can depend on the selection of smoothing parameter used to fit the background time trend. Choosing the wrong smoothing parameter yields a biased estimate of the pollution effect. The parameter cascading method may be useful in this case in that it chooses the smoothing parameters automatically by minimizing the generalized cross validation. Dominici et al. (2002) showed that the default settings in the *gam* function of the S-Plus software package (version 3.4), which implemented the back-fitting method, did not assure the convergence, and could overestimate effects of air pollution. We find the parameter cascading method converges much faster, since each level of optimization has the gradients and Hessian matrices given analytically. Moreover, Ramsay et al. (2003) showed that the *gam* function also underestimated variances of air pollution effects. We address this problem by developing the unconditional variance estimates for parameters.

The remainder of this article is organized as follows. Section 2 introduces how to estimate GSAMs using the parameter cascading method. Section 3 compares the parameter cascading method with the back-fitting method with one simulation study. Section 4 demonstrates this method with one application on finding the effect of air pollution on public health and the other application on the management of a retirement fund. Conclusions and discussion are given in Sect. 5.

2 Parameter cascading method

Let us first write Model (1) in matrix form. The functional parameters $f_i(Z_i)$, $i = 1, \ldots, P$, are estimated by linear combinations of K_i B-spline basis functions:

$$f_i(Z_i) = \sum_{k=1}^{K_i} c_{ik}\phi_{ik}(Z_i) = \mathbf{c}_i'\boldsymbol{\phi}_i(Z_i),$$

where $\mathbf{c}_i = (c_{i1}, \ldots, c_{iK_i})'$ and $\boldsymbol{\phi}_i(Z_i) = (\phi_{i1}(Z_i), \ldots, \phi_{iK_i}(Z_i))'$. Since it is not easy to choose the appropriate locations and number of knots when defining B-spline basis functions, our parameter cascading method defines a roughness penalty term to control the smoothness of $f_i(Z_i)$. Then we can choose a saturated number of basis functions. A rule of thumb is to use cubic B-spline basis functions with one knot at each location with observations.

Let $\boldsymbol{\Phi}_i$ be a $n \times K_i$ matrix with the j-th row $\boldsymbol{\phi}_i(Z_{ij})'$, then Model (1) can be written as follows:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\Phi}\mathbf{c} + \mathbf{X}\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, $\mathbf{c} = (\mathbf{c}_1', \ldots, \mathbf{c}_P')'$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_P)$ and $\mathbf{X}$ is an $n \times Q$ matrix with jk-th entry X_{kj}.

Here $\mathbf{c}$ is a vector of nuisance parameters, and $\boldsymbol{\beta}$ is a vector of structural parameters. We will introduce smoothing parameters to control the roughness of the functional parameters $f_i(Z_i)$, which is denoted as $\boldsymbol{\theta}$ as a vector. The parameter cascading method estimates $\mathbf{c}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ in three nested levels of optimization. In the first level, $\mathbf{c}$ is estimated by optimizing a criterion $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y})$, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. As a result, the conditional estimate $\widehat{\mathbf{c}}$ can be treated as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, which is denoted as $\widehat{\mathbf{c}}(\boldsymbol{\beta}, \boldsymbol{\theta})$. In the second level, a criterion $H(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})$ is optimized to obtain the estimate $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{\theta}$. Therefore, the conditional estimate $\widehat{\boldsymbol{\beta}}$ may be viewed as a function of $\boldsymbol{\theta}$. Parameter $\mathbf{c}$ is removed from the parameter space in the second level by plugging in $\widehat{\mathbf{c}}(\boldsymbol{\beta}, \boldsymbol{\theta})$. In the third level, the estimate $\widehat{\boldsymbol{\theta}}$ is acquired by optimizing another criterion $F(\boldsymbol{\theta}|\mathbf{y})$. The parameters $\mathbf{c}$ and $\boldsymbol{\beta}$ are removed from the parameter space in the third level by plugging in $\widehat{\mathbf{c}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, respectively. After obtaining the estimate $\widehat{\boldsymbol{\theta}}$, we then get the estimate $\widehat{\boldsymbol{\beta}}$ by optimizing $H(\boldsymbol{\beta}|\widehat{\boldsymbol{\theta}}, \mathbf{y})$, and the estimate $\widehat{\mathbf{c}}$ by optimizing $J(\mathbf{c}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \mathbf{y})$. We outline each optimization level in the next three subsections.

2.1 The first level: estimating nuisance parameters

The optimization criterion in the first level is written as:

$$J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) = -l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y}) + \sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 \, dZ_i, \tag{4}$$

where $l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y})$ is the log likelihood function. The second term in (4) penalizes the roughness of functional parameters, so a positive sign is used in front of it such that the optimal coefficient vector $\mathbf{c}$ can be estimated by minimizing $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$. L_i is a linear differential operator of order m:

$$L_i x(t) = \sum_{j=0}^{m-1} \alpha_j(t) \frac{d^j x(t)}{dt^j} + \frac{d^m x(t)}{dt^m},$$

where the parameter $\alpha_j(t)$ is assumed known in this article. The penalty term $\int [L_i f_i(Z_i)]^2 \, dZ_i$ can be written as a quadratic function of the coefficient vector $\mathbf{c}_i$:

$$\int [L_i f_i(Z_i)]^2 \, dZ_i = \mathbf{c}_i' \mathbf{R}_i \mathbf{c}_i,$$

where $\mathbf{R}_i = \int [L_i \boldsymbol{\phi}_i(Z_i)][L\boldsymbol{\phi}_i(Z_i)]' dZ_i$, $i = 1, \ldots, p$, is an order K_i matrix. Then, the second term in (4) can be represented in the matrix form:

$$\sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 dZ_i = \mathbf{c}'\mathbf{R}\mathbf{c},$$

where $\mathbf{c} = (\mathbf{c}_1', \ldots, \mathbf{c}_P')'$ and $\mathbf{R} = \mathrm{diag}(\lambda_1 \mathbf{R}_1, \ldots, \lambda_P \mathbf{R}_P)$. In order to attain a positive estimate for the smoothing parameter vector, we express $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_P)' = \exp(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)'$. All simulations and applications in this paper use the second derivative to define the roughness penalty term, that is, $Lx(t) = d^2x(t)/dt^2$. Ramsay and Silverman (2005) show how to obtain better estimates by defining the penalty terms with differential operators.

The estimate $\widehat{\mathbf{c}}$ is acquired by minimizing $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y})$ using the Newton-Raphson iteration method, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The first and second partial derivatives of $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y})$ with respect to $\mathbf{c}$ are given in the supplementary file.

The estimate $\widehat{\mathbf{c}}$ can be viewed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. However, there is no explicit form of this function except when observations are normally distributed. This is why least squares estimators are often used in the literature instead of likelihood functions. Fortunately, we can write out any order partial derivatives of $\widehat{\mathbf{c}}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ analytically using the Implicit Function Theorem. For example, in order to obtain the first partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\beta}$, we can take the $\boldsymbol{\beta}$-derivative on both sides of the identity $\partial J/\partial \mathbf{c}|_{\widehat{\mathbf{c}}} = 0$:

$$\frac{d}{d\boldsymbol{\beta}}\left(\frac{\partial J}{\partial \mathbf{c}}\bigg|_{\widehat{\mathbf{c}}}\right) = \frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\beta}}\bigg|_{\widehat{\mathbf{c}}} + \frac{\partial^2 J}{\partial \mathbf{c}^2}\bigg|_{\widehat{\mathbf{c}}} \frac{\partial \widehat{\mathbf{c}}}{\partial \boldsymbol{\beta}} = 0.$$

Assuming that $\partial^2 J/\partial \mathbf{c}^2|_{\widehat{\mathbf{c}}}$ is not singular, we get:

$$\frac{\partial \widehat{\mathbf{c}}}{\partial \boldsymbol{\beta}} = -\left[\frac{\partial^2 J}{\partial \mathbf{c}^2}\bigg|_{\widehat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^2 J}{\partial \mathbf{c}\partial \boldsymbol{\beta}}\bigg|_{\widehat{\mathbf{c}}}\right]. \tag{5}$$

Similarly, we can acquire other partial derivatives of $\widehat{\mathbf{c}}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ analytically, which are listed in the supplementary file. These analytical derivatives quantitatively describe the functional relationship between $\mathbf{c}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$, and are utilized in the next three subsections.

2.2 The second level: estimating structural parameters

Here we choose the negative log likelihood as the optimization criterion in the second level:

$$H(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y}) = -l(\widehat{\mathbf{c}}(\boldsymbol{\beta}, \boldsymbol{\theta}), \boldsymbol{\beta}|\mathbf{y}).$$

Notice that this criterion drops the penalty term, since the function $\widehat{\mathbf{c}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ already implicates the penalty information, and this information is passed to the log likelihood function by treating $\widehat{\mathbf{c}}$ as an implicit function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. By using this functional relationship, the coefficient vector $\widehat{\mathbf{c}}$ is eliminated from the parameter space in the second level.

We obtain the estimate $\widehat{\boldsymbol{\beta}}$ by optimizing $H(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})$ using the Newton-Raphson iteration method, conditional on any fixed value of $\boldsymbol{\theta}$. The gradient is calculated similarly as in the first level, except that we have used the chain rule to accommodate $\widehat{\mathbf{c}}$ being an implicit function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. For example, the first partial derivatives of $H(\boldsymbol{\beta}|\boldsymbol{\theta}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ is:

$$\frac{\partial H}{\partial \boldsymbol{\beta}} = -\frac{\partial l}{\partial \boldsymbol{\beta}} - \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}\right)' \frac{\partial l}{\partial \mathbf{c}}.$$

Similarly, we can get the corresponding Hessian matrix, as given in the supplementary file.

For any value of $\boldsymbol{\theta}$, we get one estimate $\widehat{\boldsymbol{\beta}}$. Therefore, the estimate $\widehat{\boldsymbol{\beta}}$ can also be treated as a function of $\boldsymbol{\theta}$. In most cases, this function is implicit, but we can again obtain analytical forms of any order partial derivatives of $\widehat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\theta}$ by the Implicit Function Theorem, as outlined in the supplementary file.

2.3 The third level: estimating complexity parameters

The smoothing parameter vector $\boldsymbol{\theta} = \ln(\boldsymbol{\lambda})$ is a complexity parameter, and controls the effective degrees of freedom of GSAMs. Let $\mathbf{D}(\mathbf{y}, \hat{\boldsymbol{\mu}})$ and $\mathbf{V}(\hat{\boldsymbol{\mu}})$ denote the deviance and variance of the model, respectively. For example, for Poisson distributed observations, $\mathbf{D}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{j=1}^{n}[y_j \ln(y_j/\hat{\mu}_j) - (y_j - \hat{\mu}_j)]$, and $\mathbf{V}(\hat{\boldsymbol{\mu}}) = \hat{\boldsymbol{\mu}}$. We apply the generalized cross validation (GCV) as the optimization criterion in the third level:

$$F(\boldsymbol{\theta}|\mathbf{y}) = \frac{n\,\mathbf{D}(\mathbf{y}, \hat{\boldsymbol{\mu}})}{[n - \alpha \mathtt{Tr}(\mathbf{A}(\boldsymbol{\theta}))]^2}, \tag{6}$$

where $\mathbf{A} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R})^{-1}\boldsymbol{\Phi}'\mathbf{W}$, $\mathbf{W} = \mathrm{diag}(w_j)$ with $w_j = [\mathbf{V}(\hat{\mu}_j)[dg(\hat{\mu}_j)/d\hat{\mu}_j]^2]^{-1}$, and $\alpha \geq 1$ is a constant. Gu and Ma (2005) suggested α in the range of 1.2–1.4 to prevent severe undersmoothing typically suffered by cross-validation methods, with little loss of general effectiveness. The results shown in the following correspond to $\alpha = 1.2$, and change little with other values of α.

The mean estimate $\hat{\boldsymbol{\mu}}$ is an explicit function of nuisance parameter $\widehat{\mathbf{c}}$ and structural parameter $\widehat{\boldsymbol{\beta}}$ as shown in (3). Since $\widehat{\mathbf{c}}$ is an implicit function of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}$, and $\widehat{\boldsymbol{\beta}}$ is an implicit function of $\boldsymbol{\theta}$, we can treat $\hat{\boldsymbol{\mu}}$, as well as $\mathbf{D}$, as an implicit function of $\boldsymbol{\theta}$. The estimate $\widehat{\boldsymbol{\theta}}$ can be acquired by minimizing $F(\boldsymbol{\theta}|\mathbf{y})$ using the Newton-Raphson iteration method. The analytic gradient and Hessian matrix are given in Appendix.

2.4 Unconditional variance estimation

The three estimates $\widehat{\mathbf{c}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are all implicit functions of the data vector $\mathbf{y}$. We develop the unconditional variance estimates, which include the uncertainty of all other parameter estimates. These variance estimates for $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{c}}$ are outlined in the following.

We approximate $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ with the first order Taylor expansion:

$$\widehat{\boldsymbol{\theta}}(\mathbf{y}) \approx \widehat{\boldsymbol{\theta}}(\boldsymbol{\mu}) + \left[\frac{d\widehat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} (\mathbf{y}-\boldsymbol{\mu}). \tag{7}$$

Taking variance on both sides of (7), we obtain the approximate variance of $\widehat{\boldsymbol{\theta}}$:

$$\mathtt{Var}[\widehat{\boldsymbol{\theta}}(\mathbf{y})] \approx \left[\frac{d\widehat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \boldsymbol{\Sigma} \left[\frac{d\widehat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]'_{\mathbf{y}=\boldsymbol{\mu}} \tag{8}$$

where $\boldsymbol{\Sigma}$ is the variance–covariance matrix for $\mathbf{y}$. The derivative $d\widehat{\boldsymbol{\theta}}/d\mathbf{y}$ can be obtained with the Implicit Function Theorem, as we do for (5). In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (8) are replaced with their estimates. Since the variance of $g(\mathbf{y})$ can be estimated by $\mathbf{r}^T\mathbf{r}/[n-\mathtt{Tr}\mathbf{A}]$ (Wahba 1985), we estimate $\boldsymbol{\Sigma}$ by the Delta method as follows:

$$\widehat{\boldsymbol{\Sigma}} = \frac{\mathbf{r}^T\mathbf{r}}{n-\mathtt{Tr}\mathbf{A}} \frac{dg^{-1}(\mathbf{y})}{d\mathbf{y}} \left(\frac{dg^{-1}(\mathbf{y})}{d\mathbf{y}}\right)',$$

where $g(\cdot)$ is the link function in Model (1) and the residual vector $\mathbf{r} = g(\mathbf{y}) - \boldsymbol{\Phi}\widehat{\mathbf{c}} - \mathbf{X}\widehat{\boldsymbol{\beta}}$. Since the observations are assumed to be independent, we set the off-diagonal entries in $\widehat{\boldsymbol{\Sigma}}$ as 0.

Similarly, the unconditional variance estimate for $\widehat{\boldsymbol{\beta}}$ is

$$\mathtt{Var}[\widehat{\boldsymbol{\beta}}(\mathbf{y})] = \left[\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \boldsymbol{\Sigma} \left[\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}}\right]'_{\mathbf{y}=\boldsymbol{\mu}}, \tag{9}$$

where the total derivative of $\widehat{\boldsymbol{\beta}}$ with respect to $\mathbf{y}$ is obtained with the chain rule:

$$\frac{d\widehat{\boldsymbol{\beta}}}{d\mathbf{y}} = \frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\widehat{\boldsymbol{\theta}}} \frac{d\widehat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\mathbf{y}}.$$

Partial derivatives $\partial\widehat{\boldsymbol{\beta}}/\partial\widehat{\boldsymbol{\theta}}$ and $\partial\widehat{\boldsymbol{\beta}}/\partial\mathbf{y}$ are given in the supplementary file. On the other hand, when we ignore the uncertainty coming from the estimate of the smoothing parameter vector $\widehat{\boldsymbol{\theta}}$, the partial derivative of $\widehat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$ is zero. The conditional variance estimate for $\widehat{\boldsymbol{\beta}}$ uses only the partial derivative of $\widehat{\boldsymbol{\beta}}$ with respect to $\mathbf{y}$ instead of the full derivative used in (9):

$$\mathtt{Var}[\widehat{\boldsymbol{\beta}}(\mathbf{y})|\hat{\boldsymbol{\lambda}}] = \left[\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \boldsymbol{\Sigma} \left[\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\mathbf{y}}\right]'_{\mathbf{y}=\boldsymbol{\mu}}.$$

The unconditional variance estimate for $\widehat{\mathbf{c}}$ is similarly obtained:

$$\mathtt{Var}[\widehat{\mathbf{c}}(\mathbf{y})] = \left[\frac{d\widehat{\mathbf{c}}}{d\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \boldsymbol{\Sigma} \left[\frac{d\widehat{\mathbf{c}}}{d\mathbf{y}}\right]'_{\mathbf{y}=\boldsymbol{\mu}}, \tag{10}$$

where the total derivative of $\widehat{\mathbf{c}}$ with respect to $\mathbf{y}$ is:

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

When we ignore the uncertainty coming from the estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$, the partial derivatives of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$ are both zero. Then the conditional variance estimate for $\widehat{\mathbf{c}}$ uses only the partial derivative of $\widehat{\mathbf{c}}$ with respect to $\mathbf{y}$ instead of the full derivative used in (10):

$$\mathtt{Var}[\widehat{\mathbf{c}}(\mathbf{y})|\widehat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}] = \left[\frac{\partial\widehat{\mathbf{c}}}{\partial\mathbf{y}}\right]_{\mathbf{y}=\boldsymbol{\mu}} \boldsymbol{\Sigma} \left[\frac{\partial\widehat{\mathbf{c}}}{\partial\mathbf{y}}\right]'_{\mathbf{y}=\boldsymbol{\mu}}. \tag{11}$$

Comparing the conditional and unconditional variance estimates, we can see that the unconditional variance estimates for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{c}}$ consider the uncertainty coming from other parameters, which addresses the underestimation problem found by Ramsay et al. (2003).

3 Simulations

The backfitting method is a popular method to estimate the generalized semiparametric additive model (Hastie and Tibshirani 1990). It is compared with the parameter cascading method based on the model

$$y_i = f(t_i) + \beta x_i + \epsilon_i, \tag{12}$$

where ϵ_i, $i = 1, \ldots, n$, are independent Gaussian errors with mean 0 and variance σ^2. The function $f(t)$ is estimated as a nonparametric function, and x_i is the covariate value. Set the true function $f(t) = \sin(2\pi t)$, and the covariate values, x_i, are generated from $\mathtt{Normal}(0, 1)$. Set $\sigma = 0.1$, which represents 15% of the signal variance, $\beta = 0.1$, and the points t_i are equally spaced in the interval $[0, 1]$,

We compare the two methods with 1000 simulations when the number of observations $n = 10, 50$, and 100. Table 1 summarizes the simulation results for both the point estimates $\hat{\beta}$ and the standard error estimates $\widehat{\mathrm{SE}}(\hat{\beta})$. It shows that the parameter cascading method has smaller biases and variances of both point estimates $\hat{\beta}$ and the standard error estimates $\widehat{\mathrm{SE}}(\hat{\beta})$ than the backfitting method.

When the data set is small ($n = 10$), both methods have some bias for the point estimates $\hat{\beta}$, which is 6% using the parameter cascading method and 11% using the backfitting method. The standard deviation of the point estimates $\hat{\beta}$ with the parameter cascading method is 49% of that with the

Table 1 The simulation results for the semiparametric additive model (12) when the number of observation $n = 10, 50, 100$

	$n = 10$			
Methods	Backfitting method		Parameter cascading	
Estimates	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$
Mean	0.0889	0.3443	0.0939	0.1850
STD	0.4323	0.1272	0.2135	0.1091
Bias ($\times 10^{-3}$)	−11.0535	−87.974	−6.1455	−28.484
	$n = 50$			
Methods	Backfitting method		Parameter cascading	
Estimates	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$
Mean	0.0996	0.1397	0.1000	0.0156
STD	0.1553	0.0208	0.0163	0.0035
Bias ($\times 10^{-3}$)	−0.4410	−15.545	−0.137	−0.7372
	$n = 100$			
Methods	Backfitting method		Parameter cascading	
Estimates	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$	$\hat{\beta}$	$\widehat{\mathrm{SE}}(\hat{\beta})$
Mean	0.0996	0.0996	0.1000	0.0104
STD	0.1023	0.0104	0.0105	0.0019
Bias ($\times 10^{-3}$)	−0.4405	−2.7422	−0.132	−0.1098

backfitting method. The backfitting method underestimates the standard error of $\hat{\beta}$ by 26%, while the parameter cascading method only underestimates the standard error of $\hat{\beta}$ by 15%. The parameter cascading method also has more stable standard error estimates, reducing 14% of the standard derivation of $\widehat{\mathrm{SE}}(\hat{\beta})$ from the backfitting method.

When the sample size is median ($n = 50$), both methods have very small bias, but the parameter cascading method reduces the bias by 69% from the backfitting method. The parameter cascading method has only 10% of the standard deviation of $\hat{\beta}$'s from the backfitting method. The backfitting method underestimates the standard error of $\hat{\beta}$ by 11%, while the parameter cascading method only 5%. The parameter cascading method also has more stable standard error estimates, with 17% of the standard deviation of $\widehat{\mathrm{SE}}(\hat{\beta})$ estimated from the backfitting method.

When the sample size increases to $n = 100$, the bias of point estimates with both methods has little change, but the standard derivation decreases 35% for both methods, compared with the results with $n = 50$. The backfitting method underestimates the standard error of $\hat{\beta}$ by 3%, while the parameter cascading method only underestimates it by 1%. The parameter cascading method also reduces 82% of the standard deviation of $\widehat{\mathrm{SE}}(\hat{\beta})$ estimated from the backfitting method. The parameter cascading method and the backfitting method take around 0.84 and 0.02 seconds for each simulation on average, respectively, in my personal laptop with the Intel Pentium Dual CPU T3400 2.16 GHz when the sample size $n = 100$. We also vary the values for σ, β and use different true function $f(t)$, and the results are consistent and therefore are not shown here to save space.

4 Applications

4.1 Effects of air pollution on public health

The parameter cascading method is used to fit Model (2) to the data displayed in Fig. 1. Using 500 cubic spline basis functions, GCV selects a smoothing parameter $\hat{\lambda} = \exp(\hat{\theta}) = 1.1 \times 10^{-4}$ with the Newton-Raphson method, corresponding to 108 effective degrees of freedom, or roughly 11 degrees of freedom per year.

The fitted linear coefficient is $\hat{\beta} = 7.3 \times 10^{-4}$, representing about a 0.07 percent increase in mortality associated with a one part-per-billion increase in ozone. The estimated standard error of $\hat{\beta}$ is 2.2×10^{-4}, yielding a 95% confidence interval of $[3.0, 11.6] \times 10^{-4}$, suggesting that the ozone effect is real. The estimated nonparametric function $\hat{f}(t)$, displayed in Fig. 2 with a 95% confidence band, shows higher mortality in the winter.

Parametric bootstrap (Efron and Tibshirani 1993) is used to assess the performance of our method. We sample 1000 Poisson data sets $\{y_j\}_{j=1}^{n}$ from the data-generating mechanism defined by the above fitted model. Figure 3 displays the estimated standard error for the nonparametric function $\hat{f}(t)$ in two years for one simulated data set. The unconditional estimate for the standard error of the nonparametric function $\hat{f}(t)$ includes the uncertainty coming from $\hat{\beta}$ and $\hat{\lambda}$, which well matches the bootstrap estimate. On the contrary, the conditional standard error estimate for $\hat{f}(t)$ underestimates the standard error by more than 10%, because it does not include the uncertainty coming from the estimates $\hat{\beta}$ and $\hat{\theta}$. The result for all ten years is similar, but hard to display in one graph.

4.2 Management of a retirement fund

Bryant and Smith (1995) introduce a managerial problem based on a real data set: Best Retirement Inc. (BRI) sells retirement plans to smaller firms with 500 or fewer employees around the United States. For a particular type of retirement plan called 401(k), the company collects the data on several attributes of the firms from the previous year. BRI would like to predict the year-end dollar amount contributed to each plan in advance in order to make internal revenue and cost projections. BRI has a sales representative who has been specifically trained to deal with 401(k) retirement plans. BRI want to know if her expertise can influence the year-end contributions. Ruppert et al. (2003) use a semiparametric additive model to answer the above questions.

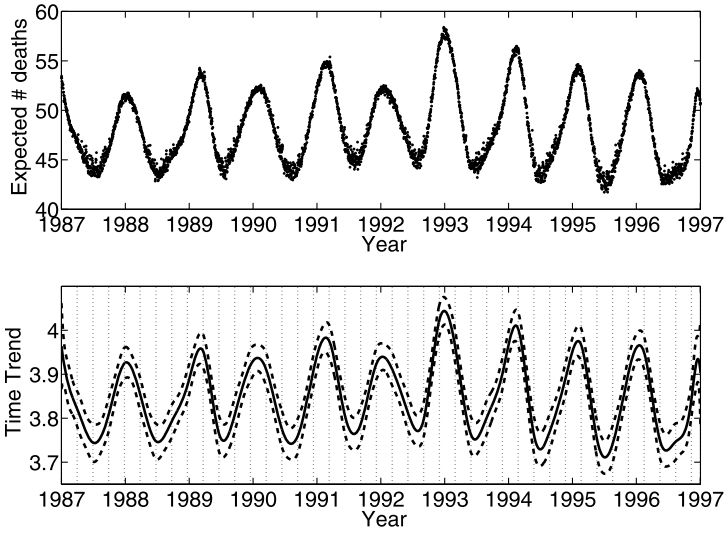

Fig. 2 The unconditional estimated expectation of daily count of non-accidental deaths from 1987 to 1996 in Toronto (*top panel*). The *bottom panel* shows the estimated functional parameter $\hat{f}(t)$ with the 95% confidence band, which is expanded by cubic B-splines with the knots indicated by the *vertical dashed lines*

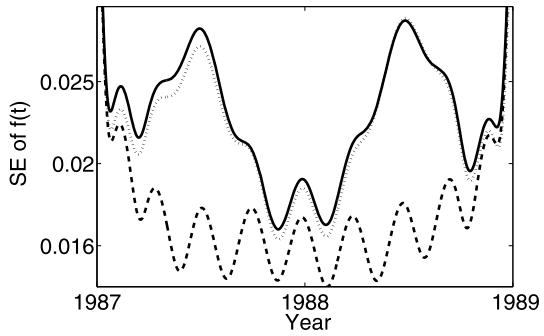

Fig. 3 The unconditional standard error estimate for the nonparametric function $\hat{f}(t)$ is estimated with the parameter cascading method (written in (10), the *solid line*). The *dashed line* is the conditional standard error estimate for the nonparametric function $\hat{f}(t)$ (written in (11)), which ignores the uncertainty coming from $\hat{\beta}$. The *dotted line* is the standard error estimate for $\hat{f}(t)$ by parametric bootstrap

The model is:

$$\mathrm{E}[\log(\text{contribution})] = \beta_1 \text{group} + \beta_2 \text{susan} + \beta_3 \text{eligible} + f(\text{salary}) \qquad (13)$$

where log(contribution) is the logarithm of the contribution to retirement plan at the end of the first year and is approximately in the normal distribution. The covariate "group" is 1 if the client has group health insurance policy, and 0, otherwise. The covariate "susan" is 1 if the plan was sold by a sales representative who has been specifically trained to deal exclusively with 401(k) plans, and 0, otherwise. The covariate "eligible" is the number of employees eligible to participate in 401(k) plans. The covariate "salary" is the average annual employee salary in dollars. The function $f(\cdot)$ is unknown and is estimated nonparametrically.

We estimate the semiparametric additive model (13) with the parameter cascading method. The function $f(\cdot)$ is approximated by a linear combination of basis functions. We choose the basis functions to be cubic B-splines with one knot at each data point. The estimated smoothing parameter $\hat{\lambda} = 0.01$. The results are shown in Table 2. The covariates "group" and "salary" have a significant effect on the contribution to retirement plan. The sales representative has a marginally significant effect on the retirement plan contribution. These results are consistent with what is found by Ruppert et al. (2003).

Table 2 Estimation results for the semiparametric additive model (13) from the real data

Parameter	β_1	β_2	β_3
Estimates	−0.25	0.33	5.5×10^{-3}
Standard errors	0.15	0.23	7×10^{-4}
t-ratio	−1.7	1.5	7.4
p-value	0.045	0.072	4×10^{-11}

5 Conclusions

The generalized semiparametric additive model is one example of a statistical model with nuisance, structural and complexity parameters, but there are many others. For example, in mixed effects models, the random effects are nuisance parameters, the fixed effects are structural parameters, and the variance parameters are complexity parameters. We develop the parameter cascading method to estimate these three kinds of parameters. Using the conditional estimates in three nested optimization levels, we obtain the functional relationships between nuisance, structural and complexity parameters, which are quantitatively described with analytical derivatives. These functional relationships may be explicit or implicit, and in the latter case the Implicit Function Theorem is exploited to obtain all the required derivatives. Al-

though these derivatives are complicated to derive, we have worked out all of them, outlined in the supplementary file. A Matlab package has also been developed to calculate them, so practitioners are not required to choose the basis functions or smoothing parameters, allowing the method to run automatically.

The gradients and Hessian matrices are derived analytically by considering these functional relationships between parameters, which decreases the computational load dramatically. We also develop the unconditional variance estimates by taking into account the functional relationships between parameters and observations, which include the uncertainty coming from other parameter estimates.

The parameter cascading method uses the implicit function theorem to obtain the derivatives, and the implicit function theorem requires the derivatives are continuous. Therefore, the parameter cascading method requires the likelihood function to be continuously differentiable. When the analytic derivatives are hard to obtain, the numeric derivative methods can be used instead.

The generalized semiparametric additive models are well estimated by the parameter cascading method for arbitrarily distributed observations, based on penalized likelihood functions. Our method is an improvement over other traditional approaches. Its benefits include fast computation, stable convergence, and unconditional variance estimates. Our results imply that the parameter cascading method is a good alternative when estimating complex statistical models with nuisance parameters.

Acknowledgements The author would like to thank Dr. T.O. Ramsay for providing the problem and air pollution data set and Dr. J.O. Ramsay for the constructive comments. This research is supported by a discovery grant from the Natural Science and Engineering Research Council of Canada (NSERC).

Appendix

In order to simplify notation, we define the matrix $\mathbf{B} = \boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R}$ and $\mathtt{dfe} = n - \alpha\mathtt{Tr}(\mathbf{A})$.

- The partial derivative of $F(\boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$:

$$\frac{\partial F(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_l} = n\left[\mathtt{dfe}\frac{\partial\mathbf{D}}{\partial\theta_l} - 2\mathbf{D}\frac{\partial\mathtt{dfe}}{\partial\theta_l}\right]\mathtt{dfe}^{-3},$$

where

$$\frac{\partial\mathbf{D}}{\partial\theta_l} = \sum_{j=1}^{n}\left(\frac{\partial\mathbf{D}}{\partial\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\theta_l}\right)$$

$$\frac{\partial\mathtt{dfe}}{\partial\theta_l} = -\alpha\mathtt{Tr}\left(\frac{\partial\mathbf{A}}{\partial\theta_l}\right)$$

$$\frac{\partial\boldsymbol{\eta}}{\partial\theta_l} = \boldsymbol{\Phi}\left[\frac{\partial\widehat{\mathbf{c}}}{\partial\theta_l} + \frac{\partial\widehat{\mathbf{c}}}{\partial\boldsymbol{\beta}}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\theta_l}\right] + \mathbf{X}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\theta_l}$$

$$\frac{\partial\mathbf{A}}{\partial\theta_l} = \boldsymbol{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_l}\boldsymbol{\Phi}'\mathbf{W} + \boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}'\frac{\partial\mathbf{W}}{\partial\theta_l}$$

$$\frac{\partial\mathbf{B}^{-1}}{\partial\theta_l} = -\mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1}$$

$$\frac{\partial\mathbf{B}}{\partial\theta_l} = \boldsymbol{\Phi}'\frac{\partial\mathbf{W}}{\partial\theta_l}\boldsymbol{\Phi} + \frac{\partial R}{\partial\theta_l}$$

$$\frac{\partial\mathbf{W}}{\partial\theta_l} = \mathrm{diag}\left(\frac{\partial w_j}{\partial\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial\eta_j}{\partial\theta_l}\right)$$

$$\frac{\partial R}{\partial\theta_l} = \mathrm{diag}(0,\ldots,0,\lambda_l\mathbf{R}_l,0,\ldots,0).$$

- The second partial derivative of $F(\boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$:

$$\frac{\partial^2 F(\boldsymbol{\lambda}|\mathbf{y})}{\partial\theta_l\partial\theta_k} = \frac{n}{\mathtt{dfe}^2}\frac{\partial^2\mathbf{D}}{\partial\theta_l\partial\theta_k} - \frac{2n\mathbf{D}}{\mathtt{dfe}^3}\frac{\partial^2\mathtt{dfe}}{\partial\theta_l\partial\theta_k} + \frac{6n\mathbf{D}}{\mathtt{dfe}^4}\frac{\partial\mathtt{dfe}}{\partial\theta_l}\frac{\partial\mathtt{dfe}}{\partial\theta_k} - \frac{2n}{\mathtt{dfe}^3}\left[\frac{\partial\mathtt{dfe}}{\partial\theta_l}\frac{\partial\mathbf{D}}{\partial\theta_k} + \frac{\partial\mathtt{dfe}}{\partial\theta_k}\frac{\partial\mathbf{D}}{\partial\theta_l}\right],$$

where

$$\frac{\partial^2\mathbf{D}}{\partial\theta_l\partial\theta_k} = \sum_{j=1}^{n}\left[\left(\frac{\partial^2\mathbf{D}}{\partial\mu_j^2}\left(\frac{\partial\mu_j}{\partial\eta_j}\right)^2 + \frac{\partial\mathbf{D}}{\partial\mu_j}\frac{\partial^2\mu_j}{\partial\eta_j^2}\right)\frac{\partial\eta_j}{\partial\theta_l}\frac{\partial\eta_j}{\partial\theta_k} + \frac{\partial\mathbf{D}}{\partial\mu_j}\frac{\partial\mu_j}{\partial\eta_j}\frac{\partial^2\eta_j}{\partial\theta_l\partial\theta_k}\right]$$

$$\frac{\partial^2\eta_j}{\partial\theta_l\partial\theta_k} = \boldsymbol{\Phi}_{(j)}\left[\frac{\partial^2\widehat{\mathbf{c}}}{\partial\theta_l\partial\theta_k} + \frac{\partial^2\widehat{\mathbf{c}}}{\partial\theta_l\partial\beta_i}\frac{\partial\hat{\beta}_i}{\partial\theta_k} + \frac{\partial^2\widehat{\mathbf{c}}}{\partial\beta\partial\beta_i}\frac{\partial\hat{\beta}_i}{\partial\theta_k}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\theta_l} + \frac{\partial^2\widehat{\mathbf{c}}}{\partial\beta\partial\theta_k}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\theta_l} + \frac{\partial\widehat{\mathbf{c}}}{\partial\boldsymbol{\beta}}\frac{\partial^2\widehat{\boldsymbol{\beta}}}{\partial\theta_l\partial\theta_k}\right] + \mathbf{X}\frac{\partial^2\widehat{\boldsymbol{\beta}}}{\partial\theta_l\partial\theta_k}$$

$$\frac{\partial^2\mathtt{dfe}}{\partial\theta_l\partial\theta_k} = -\alpha\mathtt{Tr}\left(\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial\theta_k}\right)$$

$$\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial\theta_k} = \boldsymbol{\Phi}\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial\theta_k}\boldsymbol{\Phi}'\mathbf{W} + \boldsymbol{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_l}\boldsymbol{\Phi}'\frac{\partial\mathbf{W}}{\partial\theta_k} + \boldsymbol{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_k}\boldsymbol{\Phi}'\frac{\partial\mathbf{W}}{\partial\theta_l} + \boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}'\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial\theta_k}$$

$$\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial\theta_k} = \mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_k}\mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial^2\mathbf{B}}{\partial\theta_l\partial\theta_k}\mathbf{B}^{-1} + \mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_k}\mathbf{B}^{-1}$$

$$\frac{\partial^2\mathbf{B}}{\partial\theta_l\partial\theta_k} = \boldsymbol{\Phi}'\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial\theta_k}\boldsymbol{\Phi} + \frac{\partial^2 R}{\partial\theta_l\partial\theta_k}$$

$$\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial\theta_k} = \mathrm{diag}\left(\left(\frac{\partial^2 w_j}{\partial\mu_j^2}\left(\frac{\partial\mu_j}{\partial\eta_j}\right)^2 + \frac{\partial w_j}{\partial\mu_j}\frac{\partial^2\mu_j}{\partial\eta_j^2}\right)\frac{\partial\eta_j}{\partial\theta_l}\frac{\partial\eta_j}{\partial\theta_k}\right.$$

$$+\frac{\partial w_j}{\partial \mu_j}\frac{\partial \mu_j}{\partial \eta_j}\frac{\partial^2 \eta_j}{\partial \theta_l \partial \theta_k}\Bigg)$$

$$\frac{\partial^2 R}{\partial \theta_l \partial \theta_k} = \begin{cases} \text{diag}(0,\ldots,0,\lambda_l \mathbf{R}_l,0,\ldots,0), & \text{when } l = j, \\ 0, & \text{when } l \neq j. \end{cases}$$

References

Barndorff-Nielsen, O.: On a formal for the distribution of a maximum likelihood estimator. Biometrika **70**, 343–365 (1983)

Berger, J.O., Liseo, B., Wolpert, R.L.: Integrated likelihood methods for eliminating nuisance parameters. Stat. Sci. **14**, 1–28 (1999)

Bryant, P.G., Smith, M.A.: Practical Data Analysis: Case Studies in Business. Irwin, Chicago (1995)

Cox, D., Reid, N.: Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. **49**(1), 1–39 (1987)

Cruddas, A.M., Reid, N., Cox, D.R.: A time series illustration of approximate conditional likelihood. Biometrika **76**(2), 231–237 (1989)

Dominici, F., McDermott, A., Zeger, S., Samet, J.: On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. **156**, 193–203 (2002)

Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC, New York (1993)

Fan, J., Li, R.: New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Am. Stat. Assoc. **99**, 710–723 (2004)

Gu, C., Ma, P.: Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. J. Comput. Graph. Stat. **14**, 485–504 (2005)

Härdle, W., Müller, M., Sperlich, S., Werwatz, A.: Nonparametric and Semiparametric Models. Springer, Berlin (2004)

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

He, X., Zhu, Z., Fung, W.: Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. Biometrika **89**, 579–590 (2002)

Lin, D., Ying, Z.: Semiparamtric and nonparametric regression analysis of longitudinal data. J. Am. Stat. Assoc. **96**, 103–126 (2001)

Lin, X., Carroll, R.: Semiparamtric regression for clustered data using generalized estimating equations. J. Am. Stat. Assoc. **96**, 1045–1056 (2001)

Lin, X., Carroll, R.: Semiparamtric estimation in general repeated measures problems. J. R. Stat. Soc. B **68**, 69–88 (2006)

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

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, New York (2005)

Ramsay, T.: 2005. Bias in semiparametric additive models. Tech. rep., University of Ottawa

Ramsay, T., Burnett, R., Krewski, D.: The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology **14**(1), 18–23 (2003)

Rice, J.: Convergence rates for partially splined models. Stat. Probab. Lett. **4**(44), 203–208 (1986)

Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)

Severini, T., Staniswalis, J.: Quasi-likelihood estimation in semiparametric models. J. Am. Stat. Assoc. **89**, 501–511 (1994)

Speckman, P.: Kernel smoothing in partial linear models. J. R. Stat. Soc. B **50**, 413–436 (1988)

Wahba, G.: A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Stat. **13**(4), 1378–1402 (1985)

Wolfgang, H., Liang, H., Gao, J.: Partially Linear Models. Springer, New York (2000)

Wood, S.N.: Generalized Additive Models. Chapman and Hall/CRC, New York (2006)

Zeger, S., Diggle, P.: Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. Biometrics **50**, 689–699 (1994)