Nonlinear Nonparametric Regression Models

Chunlei Ke and Yuedong Wang [*]

October 20, 2002

Abstract

Almost all of the current nonparametric regression methods such as smoothing splines, generalized additive models and varying coefficients models assume a linear relationship when nonparametric functions are regarded as parameters. In this article, we propose a general class of nonlinear nonparametric models that allow nonparametric functions to act nonlinearly. They arise in many fields as either theoretical or empirical models. Our new estimation methods are based on an extension of the Gauss-Newton method to infinite dimensional spaces and the backfitting procedure. We extend the generalized cross validation and the generalized maximum likelihood methods to estimate smoothing parameters. We establish connections between some nonlinear nonparametric models and nonlinear mixed effects models. Approximate Bayesian confidence intervals are derived for inference. We also develop a user friendly S-Plus function for fitting these models. We illustrate the methods with an application to ozone data and evaluate their finite-sample performance through simulations.

KEY WORDS: extended Gauss-Newton algorithm, nonlinear functional, nonlinear Gauss-Seidel algorithm, nonlinear mixed effects model, penalized likelihood, smoothing spline

1 Introduction

Spline smoothing is one of the most popular and powerful techniques in nonparametric regression (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994; Gu, 2002). Most research in spline smoothing, and more generally in nonparametric regression, focuses on the case where

[*] Chunlei Ke (email: cke@sjm.com) is Statistician, St. Jude Medical, Cardiac Rhythm Management Division, Sylmar, CA 91342. Yuedong Wang (email: yuedong@pstat.ucsb.edu) is Associate Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106. Yuedong Wang's research was supported by NIH Grant R01 GM58533. Address for correspondence: Yuedong Wang, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106.

observations are made directly on the unknown function f:

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \cdots, n, \tag{1}$$

where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ are observations, f is an unknown function belonging to a model space, $\boldsymbol{t} = (t_1, \cdots, t_n)^T$ are design points and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$ are random errors. In many applications observations are made indirectly, leading Wahba (1990) to consider the following general smoothing spline model

$$y_i = \mathcal{L}_i f + \epsilon_i, \quad i = 1, \cdots, n, \tag{2}$$

where f is observed through a known bounded linear functional $\mathcal{L}_i$. Model (1) is a special case of model (2) with $\mathcal{L}_i f = f(t_i)$ when the evaluation functional is bounded. Other interesting examples of $\mathcal{L}_i$ are $\mathcal{L}_i f = \int_a^b w_i(t) f(t) dt$ and $\mathcal{L}_i f = f'(t_i)$. Model (2) is very general and covers a wide range of applications. However, in some experiments such as remote sensing, f may only be observed indirectly through a nonlinear functional (O'Sullivan and Wahba, 1986). Moreover, nonlinear transformations are useful tools to relax constraints on f. For example, if f is known in prior to be positive, we may write $f = \exp(g)$ and estimate the unconstrained function g. Then g is connected to the response indirectly through a nonlinear functional.

Generalized additive models (GAM) and varying coefficients models are widely used in practice (Hastie and Tibshirani, 1990; Hastie and Tibshirani, 1993). The unknown functions act linearly in these models. Sometimes this additive linear relationship is too restrictive. For example, the true relationship may be multiplicative and/or nonlinear transformations may be needed to relax certain constraints on some additive component functions.

In this article, we consider general nonlinear nonparametric regression models (NNRMs). For a special case where unknown functions are observed through a nonlinear function of some known bounded linear functionals, we show that the penalized least squares estimates are in finite dimensional spaces and solve these estimates using Gauss-Newton and Newton-Raphson methods. For the general case, we develop an estimation procedure based on the first order approximation. The Gauss-Seidal type algorithm is used to estimate multiple functions. Smoothing parameters are estimated by the extended generalized cross-validation and generalized maximum likelihood methods. Approximate Bayesian confidence intervals will be derived for inference.

In Section 2, we introduce NNRMs. Estimation methods are proposed in Section 3. Connections between some special NNRMs and nonlinear mixed effects models are established in Section 4. In Section 5, approximate Bayesian confidence intervals for the unknown functions are constructed based on linear approximation. In section 6, we briefly introduce a generic S-Plus program for fitting NNRMs. Application to a real data set is described in Section 7. In Section 8, we conduct several simulations to evaluate our estimation and inference procedures. This article ends with discussions of topics for future research.

2 The Model

For simplicity, throughout this paper we refer to models (1) and (2) as linear smoothing spline models to indicate the linear relationship between the expected responses and nonparametric functions when nonparametric functions are regraded as parameters. They should not be confused with the linear polynomial splines.

This article aims to extend the linear smoothing spline models (2) and GAM-type models to the *nonlinear* case. Specifically, we define a general NNRM as

$$y_i = \mathcal{N}_i(g_1, \cdots, g_r) + \epsilon_i, \quad i = 1, \cdots, n, \tag{3}$$

where $\mathcal{N}_i$ is a known nonlinear functional, $\boldsymbol{g} = (g_1, \cdots, g_r)$ are unknown functions, and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$ are random errors with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. As in nonlinear regression models, g_k's are regarded as parameters. Each g_k may be a multivariate function. Throughout this article, we assume that $g_k \in \mathcal{H}_k$, where $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) on a domain $\mathcal{T}_k$. The domains $\mathcal{T}_k$'s are arbitrary sets which may be the same or different. We further assume that $\mathcal{H}_k = \mathcal{H}_{k0} \oplus \mathcal{H}_{k1}$, where $\mathcal{H}_{k0} = span\{\phi_{k1}(t), \cdots, \phi_{km_k}(t)\}$ and $\mathcal{H}_{k1}$ is a RKHS with reproducing kernel (RK) $R_{k1}(s,t)$. For example, the well-known cubic spline corresponds to

$$\mathcal{H}_k = W_2([0,1]) = \left\{ f : f \text{ and } f' \text{ are absolute continuous, } \int_0^1 (f''(t))^2 dt < \infty \right\}, \tag{4}$$

$\mathcal{H}_{k0} = \text{span}\{1, t - 0.5\}$, $R_{k1}(s,t) = k_2(s)k_2(t) - k_4(s-t)$ where $k_1(x) = x - 0.5$, $k_2(x) = \frac{1}{2}(k_1^2(x) - \frac{1}{12})$ and $k_4(x) = \frac{1}{24}(k_1^4(x) - \frac{1}{2}k_1^2(x) + \frac{7}{240})$. See Wahba (1990), Gu (2002) and Aronszajn (1950) for details on RKHS.

Obviously if $\mathcal{N}_i$ is linear in all g_k's, the NNRM reduces to an additive model. The following examples illustrate potential value of the NNRMs. Application to a real data set is presented in Section 7.

Example 1: Remote sensing (O'Sullivan and Wahba, 1985). The satellite up-welling radiance measurements R_v are related to the underlying atmospheric temperature distribution T through

$$R_v(T) = B_v(T(x_s))\tau_v(x_s) - \int_{x_0}^{x_s} B_v(T(x))\tau_v'(x)dx,$$

where x is some monotone transformation of pressure p (Meteorologists often use *kappa* units, $x(p) = p^{5/8}$); x_0 and x_s are x values at the surface and top of the atmosphere; $\tau_v(x)$ is the transmittance of the atmosphere above x at wavenumber v; and $B_v(T)$ is the Plank's function, $B_v(T) = c_1 v^3/[exp(c_2 v/T) - 1]$, with known constants c_1 and c_2. T, as a function of x, needs to be recovered from noisy observations of $R_v(T)$. Obviously $R_v(T)$ is nonlinear in T. Other examples such as reservoir modeling and three dimensional atmospheric temperature distribution from satellite-observed radiances can be found in Wahba (1987) and O'Sullivan (1986).

Example 2: Positive linear and nonlinear inverse problem (Wahba, 1990; Vardi and Lee, 1993). Linear and nonlinear inverse problems are ubiquitous in science and engineering. For example, suppose that observations are generated by the Fredholm's integral equation of the first kind

$$y_i = \int K(t_i, s) f(s) ds + \epsilon_i, \quad i = 1, \cdots, n, \tag{5}$$

where K is a known impulse response function, ϵ_i's are measurement errors which sometimes are assumed to be zero. The goal is to recover f through observations. Often f is known to be positive (Vardi and Lee 1993). Writing $f = \exp(g)$, then model (5) becomes a NNRM. Nonlinear inverse problems are an area of active research in applied mathematics. More examples can be found in Engl et al (1996). Our methods in this paper provide an alternative approach for these difficult problems with data driven choices of regularization parameters.

Example 3: Extensions of additive models. Recall that an additive model assumes that (Hastie and Tibshirani, 1990)

$$f(x_1, \cdots, x_r) = \alpha + f_1(x_1) + \cdots + f_r(x_r). \tag{6}$$

A simple extension is to assume nonlinear transformations for some or all components:

$$f(x_1, \cdots, x_r) = \alpha + \mathcal{N}^1 g_1 + \cdots + \mathcal{N}^r g_r,$$

where $\mathcal{N}^i$ are known linear or nonlinear operators. For example, if f_1 in (6) is known to be strictly increasing, then $f_1'(t) > 0$. Let $f_1'(t) = \exp(g_1(t))$. We can re-express f_1 as $f_1(t) = f_1(0) + \int_0^t \exp(g_1(s)) ds$. The constant $f_1(0)$ is absorbed by α. Therefore we have $\mathcal{N}^1 g_1 = \int_0^{x_1} \exp(g_1(t)) dt$.

Example 4: L-spline (Wahba, 1990; Gu, 2002). An L-spline, based on a linear differential operator $\mathcal{L} = D^m + \sum_{j=1}^{m-1} a_j(t) D^j$ where D^j denotes the jth derivative, provides powerful tools to test whether certain parametric models are appropriate. An L-spline can also reduce bias of a spline estimate if the true function is in the kernel (null) space of $\mathcal{L}$ (Heckman and Ramsay, 2000). Note that L-splines are defined for *linear* differential operators. Sometimes prior knowledge or physical law suggests that the true function is close to the solution of a *nonlinear* differential equation. In fact, Ramsay (1998)'s penalty on the relative curvature, $D^2 f / Df$, is a nonlinear operator. It is sometimes possible to find a (nonlinear) transformation of f that changes the original difficult problem into one with known answers. This approach is well known in the ODE literature as substitution (Sachdev, 1991; Miller, 1991). For example, the Riccati equation, $Df = h_0(t) + h_1(t) f + h_2(t) f^2$, is transformed into a linear differential equation $\mathcal{L} g = 0$ with $\mathcal{L} = D^2 - (h_2'(t)/h_2(t) + h_1(t)) D + h_0(t) h_2(t)$ and $f = -g'(t)/(h_2(t) g(t))$. Even when $\mathcal{L}$ is linear, closed form reproducing kernels are available for a few simple operators only. A nonlinear transformation can often reduce a difficult problem into a much simpler one.

For example, we can use the operator $\mathcal{L} = D + e^t/(1+e^t)^2$ if the function to be estimated is close to an inverse logistic function: $e^t/(1+e^t)$. Clearly the inverse logistic function is the solution of $\mathcal{L}f = 0$. We can transform this complicated problem into a simple one using either one of the following two transformations: (a) let $f = 1/g$, and model g using the exponential spline with $\mathcal{L} = D^2 + D$ (Gu, 2002; Wang and Ke, 2002). Then the basis for the null space is $1 + e^{-t}$ and the RK for $\mathcal{H}_1$ has a simple form; (b) let $f = e^g/(1+e^g)$, and model g using a cubic spline. The combination of L-splines with nonlinear transformations provides powerful tools to deal with complicated problems.

There has not been much research done for the general NNRMs. O'Sullivan and Wahba (1985) and O'Sullivan (1986) considered nonlinear spline models with a single unknown function. They proposed to approximate g by a finite collection of basis functions. In this article, a general NNRM with multiple functions is considered. We stress their applications as a tool to build empirical models and to reduce constraints. We propose different estimation and inference procedures which are relatively easy to implement. We also develop user friendly S-Plus functions for fitting NNRMs.

3 Estimation

We estimate $\boldsymbol{g}$ by minimizing the following penalized least squares

$$PLS = \sum_{i=1}^{n}(y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 + n\lambda \sum_{k=1}^{r} \theta_k ||P_{k1} g_k||^2, \tag{7}$$

where the first part measures the goodness-of-fit, P_{k1} is the projection operator onto the subspace $\mathcal{H}_{k1}$ in $\mathcal{H}_k$, and λ and θ_k's are smoothing parameters which balance the trade-off between the goodness-of-fit and the penalty terms. Choices of the penalty form depend on specific applications. For instance, a cubic spline on $[0,1]$ corresponds to $||P_1 g||^2 = \int_0^1 (g''(s))^2 ds$. If g_k is multivariate, then $||P_{k1} g_k||^2$ may include more than one penalty term.

For $r = 1$, O'Sullivan (1990) showed that under certain regularity conditions, the PLS criterion in (7) has a unique minimizer. Throughout this paper we assume that a solution to (7) exists. For linear operators, it is known that the solutions to the PLS fall into a finite dimensional space. Unfortunately, this classical result no longer holds in general for NNRMs (but see a special case below). Therefore, certain approximation may be necessary. O'Sullivan and Wahba (1985) approximated g by a finite series. Further research is necessary in evaluating different bases. The choice of bases may be especially difficult for multivariate functions.

In the following we first consider a special case. Then we propose a general iterative estimation procedure by approximating the nonlinear functional by using a first order Taylor expansion. This procedure is a natural extension of the Gauss-Newton method for nonlinear regression models (Bates and Watts, 1988). By combining this procedure with a nonlinear Gauss-Seidel method, we propose an algorithm for estimating multivariate functions.

3.1 A Special Case

Consider the special case

$$\mathcal{N}_i(g_1, \cdots, g_r) = \eta_i(\mathcal{L}_{1i}g_1, \cdots, \mathcal{L}_{ri}g_r), \quad i = 1, \cdots, n, \tag{8}$$

for some known nonlinear function η_i and bounded linear operators $\mathcal{L}_{1i}, \cdots, \mathcal{L}_{ri}$. This is to assume that $\mathcal{N}_i(g_1, \cdots, g_r)$ depends on g_k through $\mathcal{L}_{ki}g_k$ only. This special case covers many interesting applications. An important exception is $\mathcal{N}g = \int_0^t \exp(g(s))ds$.

Let $\delta_{ki}(s) = \mathcal{L}_{ki}R_{k1}(s, \cdot)$ and $\xi_{ki} = P_{1k}\delta_{ki}$. Extending O'Sullivan et al (1986), in Appendix A we prove

Theorem 1. For the special case (8), the solution to (7) can be represented as

$$\hat{g}_k(t) = \sum_{i=1}^{m_k} d_{ki}\phi_{ki}(t) + \sum_{j=1}^{n} \theta_k c_{kj}\xi_{kj}(t), \quad k = 1, \cdots, r. \tag{9}$$

Remark: Theorem 1 also holds when the least squares term in (7) is replaced by the negative log-likelihood. Thus it covers the cases when random errors are correlated or when observations are non-Gaussian.

Let $\boldsymbol{c}_k = (c_{k1}, \cdots, c_{kn})^T$, $\boldsymbol{c} = (\boldsymbol{c}_1^T, \cdots, \boldsymbol{c}_r^T)^T$, $\boldsymbol{d}_k = (d_{k1}, \cdots, d_{km_k})^T$ and $\boldsymbol{d} = (\boldsymbol{d}_1^T, \cdots, \boldsymbol{d}_r^T)^T$. Based on Theorem 1, the penalized least squares (7) becomes

$$PLS(\boldsymbol{c}, \boldsymbol{d}) = \sum_{i=1}^{n}(y_i - \eta_i(\mathcal{L}_{1i}\hat{g}_1, \cdots, \mathcal{L}_{ri}\hat{g}_r))^2 + n\lambda \sum_{k=1}^{r} \theta_k \boldsymbol{c}_k^T \boldsymbol{\Sigma}_k \boldsymbol{c}_k, \tag{10}$$

where $\boldsymbol{\Sigma}_k = \{< \xi_{ki}, \xi_{kj} >\}$. We need to find minimizers $\boldsymbol{c}$ and $\boldsymbol{d}$. Standard nonlinear optimization procedures such as the Gauss-Newton and Newton-Raphson methods can be employed to solve (10). `RKPACK` (Gu, 1989) can be used to update $\boldsymbol{c}$ and $\boldsymbol{d}$ at each iteration of these procedures. These procedures are described in Appendix B. We defer methods for choosing smoothing parameters to Section 3.4.

3.2 Extended Gauss-Newton Algorithm

We restrict our attention to $r = 1$ in this section. The case $r > 1$ will be discussed in the next section. For simplicity of notation, we drop the subscript k whenever $r = 1$ throughout this paper. Assume that the Fréchet differential of $\mathcal{N}_i$ with respect to g evaluated at g_- exists and is bounded. Define $\mathcal{D}_i = \partial\mathcal{N}_i/\partial g|_{g=g_-}$, which is a linear and continuous operator (Lusternik and Sobolev, 1974). We approximate $\mathcal{N}_i g$ by its first order Taylor expansion at g_- (Lusternik and Sobolev, 1974):

$$\mathcal{N}_i g \approx \mathcal{N}_i g_- + \mathcal{D}_i(g - g_-). \tag{11}$$

Then (3) is approximated by

$$\tilde{y}_i = \mathcal{D}_i g + \epsilon_i, \quad i = 1, \cdots, n, \tag{12}$$

where $\tilde{y}_i = y_i - \mathcal{N}_i g_- + \mathcal{D}_i g_-$. We minimize

$$\sum_{i=1}^{n} (\tilde{y}_i - \mathcal{D}_i g)^2 + n\lambda ||P_1 g||^2 \tag{13}$$

to get a new estimate of g. Since $\mathcal{D}_i$ is a linear and bounded functional, the solution to (13) has the form

$$\tilde{g}(t) = \sum_{i=1}^{m} \tilde{d}_i \phi_i(t) + \sum_{j=1}^{n} \tilde{c}_j \tilde{\xi}_j(t), \tag{14}$$

where $\tilde{\xi}_j(t) = \mathcal{D}_{j(\cdot)} R_1(\cdot, t)$. $\tilde{g}(t)$ can be calculated using existing software such as `RKPACK` (Gu, 1989). An iterative algorithm can then be formed with the convergent solution as the final estimate of g. Note that the linear functionals $\mathcal{D}_i$'s depend on the current estimate. Thus representers $\tilde{\xi}_j$'s change along iterations.

This algorithm could be seen as an extension of the Gauss-Newton method to infinite dimensional spaces. As in nonlinear regression, the performance of this algorithm depends largely on the curvature of the nonlinear functional. Simulations in Section 5 indicate that the algorithm works well and converges quickly for commonly used nonlinear functionals. O'Sullivan (1986) used this linearization to compute the averaging kernel for the ill-posed nonlinear inverse problem.

For the special case (8), the following result confirms that this algorithm is indeed a natural extension of the classical Gauss-Newton method.

Theorem 2. For the special case (8), the extended Gauss-Newton (EGN) algorithm is equivalent to solving (10) using the classical Gauss-Newton (CGN) method.

[Proof] See Appendix C.

3.3 Estimation of Multiple Functions

When $r > 1$, the linearization procedure may be applied to all functions simultaneously. However, it may be computationally intensive when n or r are large. We propose a Gauss-Seidel-type algorithm to estimate functions iteratively one at a time (Hastie and Tibshirani, 1990). This suggests the following algorithm.

Algorithm

(1) Initialize: $g_i = g_i^0, \quad i = 1, \cdots, r$;

(2) Cycle: for $k = 1, \cdots, r, 1, \cdots, r, \cdots$, conditional on the current estimates of $g_1, \cdots, g_{k-1}, g_{k+1}, \cdots, g_r$, g_k is updated as the minimizer of

$$\sum_{i=1}^{n} (y_i - \mathcal{N}_i(g_1, \cdots, g_{k-1}, g_k, g_{k+1}, \cdots, g_r))^2 + n\lambda ||P_{k1} g_k||^2; \tag{15}$$

(3) Continue step (2) until individual functions do not change.
Step (2) can be computed using procedures discussed in Sections 3.1 or 3.2. An inner iteration is needed at this step when $\mathcal{N}_i$ is nonlinear in g_k. From our experience, the convergence of this inner iteration is not necessary and a small number of iterations is usually good enough. Current estimate of g_k is used as initial values for the inner iteration. As in Jiang (2000), the whole procedure is referred to as the Nonlinear Gauss-Seidel algorithm (NGS). For GAM, this algorithm was called backfitting (Hastie and Tibshirani, 1990).

For the special case (8), step (2) updates $\boldsymbol{c}_k$ and $\boldsymbol{d}_k$ by minimizing

$$\sum_{i=1}^{n}(y_i - \eta_i(g_1, \cdots, g_{k-1}, \sum_{u=1}^{m_k} d_{ku}\mathcal{L}_{ki}\phi_{ku} + \sum_{v=1}^{n}\theta_k c_{kv}\mathcal{L}_{ki}\xi_{kv}, g_{k+1}, \cdots, g_r))^2 + n\lambda\theta_k \boldsymbol{c}_k^T \boldsymbol{\Sigma}_k \boldsymbol{c}_k. \quad (16)$$

Suppose that the solution to (10) exists and is unique. Then by the global convergence theorem in nonlinear programming, the NGS algorithm converges to the unique solution for fixed λ and $\boldsymbol{\theta}$ (Jiang, 2000). Conditions for the existence of a solution to (10) and its uniqueness are the same as those in nonlinear regression. A sufficient condition is that $PLS(\boldsymbol{c}, \boldsymbol{d})$ is convex in $\boldsymbol{c}$ and $\boldsymbol{d}$. Existence and uniqueness conditions for (7) can be found in O'Sullivan (1990).

3.4 Choosing the Smoothing Parameters

As usual, the smoothing parameters, $\lambda/\theta_1, \cdots, \lambda/\theta_r$, are critical to the performance of spline estimates. Furthermore, from our experience, lack of convergence is usually caused by poor choices of smoothing parameters. In this section we extend the cross validation and generalized maximum likelihood methods to select the smoothing parameters in NNRMs (O'Sullivan and Wahba, 1985; Ramsay, 1998; Wahba, 1990).

The ordinary leave-out-one cross validation method selects $\lambda/\theta_1, \cdots, \lambda/\theta_r$ as minimizers of the following score:

$$OCV(\lambda, \boldsymbol{\theta}) = \frac{1}{n}\sum_{v=1}^{n}(y_v - \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]}))^2,$$

where $g_1^{[v]}, \cdots, g_r^{[v]}$ are minimizers of

$$\sum_{i \neq v}(y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 + n\lambda\sum_{k=1}^{r}\theta_k ||P_{k1}g_k||^2. \quad (17)$$

As for the case of linear spline models (2), the following "leaving-out-one" lemma still holds.
Lemma 1. Let $g_1^{[v]}, \cdots, g_r^{[v]}$ be the minimizers of (17). For fixed v and z, let $\boldsymbol{h}[v, z]$ be the vector of functions that minimizes

$$(z - \mathcal{N}_v(g_1, \cdots, g_r))^2 + \sum_{i \neq v}(y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 + n\lambda\sum_{k=1}^{r}\theta_k ||P_{1k}g_k||^2.$$

Then $\boldsymbol{h}[v, \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]})] = (g_1^{[v]}, \cdots, g_r^{[v]})$.
[Proof] See page 52 of Wahba (1990).

Let

$$\Delta_v(\lambda, \boldsymbol{\theta}) = \frac{\mathcal{N}_v(g_1, \cdots, g_r) - \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]})}{y_v - \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]})}.$$

Then we have

$$y_v - \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]}) = \frac{y_v - \mathcal{N}_v(g_1, \cdots, g_r)}{1 - \Delta_v(\lambda, \boldsymbol{\theta})}.$$

Let

$$a_{ij} = \partial \mathcal{N}_i(\boldsymbol{h}[i, y_j])/\partial y_j = \sum_{u=1}^{r} \frac{\partial \mathcal{N}_i(g_1, \cdots, g_r)}{\partial g_u} \frac{\partial g_u}{\partial y_j}. \tag{18}$$

Let $\boldsymbol{A} = (a_{ij})_{i,j=1}^n$ and $y_v^{[v]} = \mathcal{N}_v(g_1^{[v]}, \cdots, g_r^{[v]})$. From Lemma 1, we have

$$\Delta_v(\lambda, \boldsymbol{\theta}) = \frac{\mathcal{N}_v(\boldsymbol{h}[v, y_v]) - \mathcal{N}_v(\boldsymbol{h}[v, y_v^{[v]}])}{y_v - y_v^{[v]}} \approx \frac{\partial \mathcal{N}_v(\boldsymbol{h}[v, y_v])}{\partial y_v} = a_{vv}.$$

Thus we can approximate the OCV criterion by

$$OCV(\lambda, \boldsymbol{\theta}) \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 / (1 - a_{ii})^2.$$

Similar to the GCV method for the linear case, we can further approximate the OCV score by the following GCV score

$$V(\lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 / \left[\frac{1}{n} tr(\boldsymbol{I} - \boldsymbol{A})\right]^2. \tag{19}$$

Correspondingly, we define a GML score as

$$M(\lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 / \left[\det{}^+(I - \boldsymbol{A})\right]^{1/(n - \sum_{k=1}^{r} m_k)}, \tag{20}$$

where $\det^+$ is the product of nonzero eigenvalues. Note that the generalized degrees of freedom defined in Ye (1998) equals $\mathrm{E}(tr\boldsymbol{A})$.

Since the estimates of $\boldsymbol{g}$ depend on $\boldsymbol{y}$ nonlinearly and $\partial \mathcal{N}_i(g_1, \cdots, g_r)/\partial g_k$ depends on the true functions, it is impossible to calculate a_{ij} from (18). Thus it is not feasible to minimize the GCV or GML score directly. One approach would be to approximate $\partial \mathcal{N}_i(g_1, \cdots, g_r)/\partial g_k$ by $\partial \mathcal{N}_i(\hat{g}_1, \cdots, g_k, \cdots, \hat{g}_r)/\partial g_k|_{g_k = \hat{g}_k}$, and $\partial g_k / \partial y_v$ by its corresponding component in the approximate linear model based on $\hat{g}_1, \cdots, \hat{g}_r$. This leads to approximating $\boldsymbol{A}$ by $\sum_{k=1}^{r} \boldsymbol{A}_k$, where $\boldsymbol{A}_k$ is the hat matrix for g_k based on the approximate linear model at convergence with other functions fixed at their final estimates. Minimizers of the approximated $V(\lambda, \boldsymbol{\theta})$ and $M(\lambda, \boldsymbol{\theta})$

can be found by a grid search (Xiang and Wahba, 1996). This approach is computationally intensive.

Another approach is to replace g_i's in (18) by their current estimates. This suggests to estimate λ and $\boldsymbol{\theta}$ at each iteration. Specifically, at the l-th iteration, select the optimal λ^l for the linear model (12) by the standard GCV or GML method. See Appendix B for more detail. This approach was used to estimate smoothing parameters with non-Gaussian data (Gu, 1992; Wahba et al, 1995). We call this approach as the iterative GCV and the iterative GML methods. This approach is easy to implement. However, it does not guarantee convergence (Gu, 1992). See Section 4 for a justification of the iterative GML method. From our experience, convergence is achieved in most cases. When the backfitting algorithm is used for the case of multiple functions, smoothing parameters are estimated for each k in Step (2).

We estimate σ^2 by

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \mathcal{N}_i(g_1, \cdots, g_r))^2 / (n - \sum_{k=1}^{r} \mathrm{tr}(\boldsymbol{A}_k)).$$

4 Connection to Nonlinear Mixed Effects Models

It is well known that the general spline model (2) has connections with linear mixed effects models (Wang, 1998a). These connections allow one field to borrow methods from the other. In this section, the connections between nonlinear mixed effects models (NLMMs) and NNRMs are established. For simplicity, the case of a single function ($r = 1$) is considered. It is trivial to extend to the case of multiple functions.

We first consider the special case $\mathcal{N}_i g = \eta_i(\mathcal{L}_i g)$. Then the solution to (7) has the form (9). Let $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_n)^T$, $\eta(\boldsymbol{\psi}) = (\eta_1(\psi_1), \cdots, \eta_n(\psi_n))^T$, $\boldsymbol{T} = \{\mathcal{L}_i \phi_j\}_{i=1}^{n}{}_{j=1}^{m}$ and $\boldsymbol{\Sigma} = \{\mathcal{L}_i \mathcal{L}_j R_1\}_{i,j=1}^{n}$. Consider the following NLMM

$$\begin{aligned} \boldsymbol{y} &= \eta(\boldsymbol{\psi}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}), \\ \boldsymbol{\psi} &= \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c}, \qquad \boldsymbol{c} \sim N(0, \sigma^2 \boldsymbol{\Sigma}^+ / n\lambda), \end{aligned} \tag{21}$$

where $\boldsymbol{d}$ are fixed effects, $\boldsymbol{c}$ are random effects independent of the random errors $\boldsymbol{\epsilon}$, and $\boldsymbol{\Sigma}^+$ is the Moore-Penrose inverse of $\boldsymbol{\Sigma}$. It is common practice to estimate $\boldsymbol{d}$ and $\boldsymbol{c}$ as minimizers of the following joint negative log-likelihood (Lindstrom and Bates, 1990; Lee and Nelder, 1996)

$$||\boldsymbol{y} - \eta(\boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c})||^2 + n\lambda \boldsymbol{c}^T \boldsymbol{\Sigma} \boldsymbol{c},$$

which is the same as (10). This connection to NLMM suggests alternative ways to fit NNRMs. For example, the two-step procedure in Lindstrom and Bates (1990) can be used. Denote $\boldsymbol{c}_-$ and $\boldsymbol{d}_-$ as the current estimates. At the LME step, Lindstrom and Bates (1990) approximated model (21) by the following linear mixed effects model

$$\boldsymbol{w} = \boldsymbol{X}\boldsymbol{d} + \boldsymbol{Z}\boldsymbol{c} + \boldsymbol{\epsilon}, \tag{22}$$

where $\boldsymbol{X} = \partial\eta(\boldsymbol{Td} + \boldsymbol{\Sigma c})/\partial \boldsymbol{d}|_{\boldsymbol{c}=\boldsymbol{c}_-,\boldsymbol{d}=\boldsymbol{d}_-}$, $\boldsymbol{Z} = \partial\eta(\boldsymbol{Td} + \boldsymbol{\Sigma c})/\partial c|_{\boldsymbol{c}=\boldsymbol{c}_-,\boldsymbol{d}=\boldsymbol{d}_-}$, and $\boldsymbol{w} = \boldsymbol{y} - \eta(\boldsymbol{Td}_- + \boldsymbol{\Sigma c}_-) + \boldsymbol{Xd}_- + \boldsymbol{Zc}_-$. It is not difficult to see that $\boldsymbol{w} = \check{\boldsymbol{y}}$, $\boldsymbol{X} = \boldsymbol{VT}$ and $\boldsymbol{Z} = \boldsymbol{V\Sigma}$, where $\check{\boldsymbol{y}}$ and $\boldsymbol{V}$ are defined in Appendix B. Thus it is easy to show that the REML estimate of λ based on model (22) is the same as the GML estimate for the linear spline model corresponding to equation (27) (Wang, 1998a). Therefore, the REML method for estimating λ in model (21) is the same as the iterative GML method. This provides a justification for the iterative GML method proposed in Section 3.4.

Such a connection is impossible when the solution to (7) is not in a finite dimensional space because model (21) is a finite dimensional model. However, if a finite dimensional approximation is used as the solution to (7), similar connections can be established.

5 Inference

Bayesian confidence intervals are often constructed for smoothing spline estimates for the purpose of inference (Wahba, 1990). In this section, approximate Bayesian confidence intervals are constructed for NNRMs based on linearization at convergence.

For simplicity, we again consider the case $r = 1$. Results in this section can be easily extended to the case $r > 1$. At convergence, the extended Gauss-Newton method approximates the original model by

$$\tilde{y}_i^* = \mathcal{D}_i^* g + \epsilon_i, \quad i = 1, \cdots, n,$$

where $\mathcal{D}_i^* = \partial\mathcal{N}_i/\partial g|_{g=\hat{g}}$, $\hat{g}$ is the estimate at convergence, and $\tilde{y}_i^* = y_i - \mathcal{N}_i\hat{g} + \mathcal{D}_i^*\hat{g}$. Assume a prior distribution for g as

$$G(t) = \sum_{i=1}^{m} d_i\phi_i(t) + \tau^{1/2}Z(t),$$

where $d_i \overset{iid}{\sim} \mathcal{N}(0, a)$ and $Z(t)$ is a mean zero Gaussian process with $\mathrm{Cov}(Z(s), Z(t)) = R_1(s, t)$. Now consider $\hat{g}$ as fixed, and assume that observations are generated from the following model

$$\tilde{y}_i^* = \mathcal{D}_i^* G + \epsilon_i, \quad i = 1, \cdots, n. \tag{23}$$

Since $\mathcal{D}_i^*$'s are linear operators, the posterior mean of the Bayesian model (23) equals $\hat{g}(t)$ (Wahba, 1990). Posterior variances and Bayesian confidence intervals can be calculated as in Wahba (1990). For $r > 1$, Bayesian confidence intervals can be constructed for each component as in Gu and Wahba (1993).

Since these intervals are based on the linear model (23), their performances depend largely on the accuracy of the linear approximation. When curvature of $\mathcal{N}_i$ respect to g is high, modification is necessary to improve coverage (Bates and Watts, 1981; Hamilton, Watts and Bates, 1982).

Another approach is to use the bootstrap method to construct confidence intervals. A bootstrap sample $\boldsymbol{y}^* = (y_1^*, \cdots, y_n^*)^T$ is generated through

$$y_i^* = \mathcal{N}_i(\hat{g}_1, \cdots, \hat{g}_r) + \epsilon_i^*, \quad i = 1, \cdots n,$$

where $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \cdots, \epsilon_n^*)^T$ are random samples from either $N(0, \hat{\sigma}^2)$ or residuals. New estimates $\boldsymbol{g}^*$ are then computed. This procedure is repeated for a number of times and bootstrap confidence intervals are constructed. Wang and Wahba (1995) studied the bootstrap confidence intervals for the linear smoothing spline models. They found that the bootstrap intervals are comparable to Bayesian confidence intervals and have the same "across-the-function" coverage property. However, they are computationally intensive.

6 S-Plus Function `nnr`

We now briefly describe a generic user-friendly S-Plus function, `nnr`, for fitting NNRMs. A typical call is

```
        nnr(formula, func, start, spar, data)
```

where `formula` is a two-sided formula specifying the response on the left of a ~ operator, and the model (3) on the right with nonparametric functions g_k's treated as parameters; `func` is a list with each component specifying bases $\phi_{k1}, \cdots, \phi_{km_k}$ for the null space $\mathcal{H}_{k0}$ and RK R_{k1} for $\mathcal{H}_{k1}$; `start` specifies initial values for $\boldsymbol{g}$; and `spar` specifies a method for choosing the smoothing parameters.

Both Gauss-Newton and Newton-Raphson methods are implemented in `nnr`. They are combined with backfitting for estimating multiple functions. Smoothing parameters are estimated by the iterative GCV or the iterative GML method. Supporting functions include `nnr.control`, `predict.nnr` and `intervals.nnr`. Approximate Bayesian confidence intervals can be calculated by a call to the `intervals.nnr` function. An example will be given in Section 7. `nnr` is one function in the ASSIST package which can be downloaded from `www.pstat.ucsb.edu/faculty/yuedong/research`. More detail and examples can be found in the manual of the ASSIST package which also can be found at the above website.

7 Application to Ozone Data

For illustration, we apply our methods and software to a real data set in Andrews and Herzberg (1985). Monthly mean ozone `thickness` (Dobson units) in Arosa, Switzerland was recorded from 1926-1971. We are interested in investigating changes in `thickness` over months and years. For simplicity, we scale both `year` and `month` variables into $[0, 1]$ and refer to these new variables as `csyear` and `csmonth`. It is known that ozone `thinkness` is a periodic function of month. However, the form of this periodic pattern may be unknown. Furthermore, the mean and amplitude may change over years. Thus, we consider the following NNRM

$$\texttt{thickness(csyear, csmonth)} = g_1(\texttt{csyear}) + \exp(g_2(\texttt{csyear})) \times g_3(\texttt{csmonth}) + \epsilon, \qquad (24)$$

where $g_1(\texttt{csyear})$ is the mean ozone thickness in csyear, $\exp(g_2(\texttt{csyear}))$ is the amplitude in csyear, and g_3 is the periodic pattern within a year. Model (24) is used to investigate how the mean and amplitude change over years. It generalizes GAM by allowing multiplicative components. Regarding g_1 and $\exp(g_2)$ as coefficients of the periodic pattern, model (24) also generalizes the varying coefficients models by allowing a nonparametric transformation (or model) for the covariate. Note that the exponential transformation is used to guarantee positive amplitude.

To make model (24) identifiable, we use the following side conditions: (a) $\int_0^1 g_2(t)dt = 0$, which eliminates the multiplicative constant making g_2 identifiable with g_3; and (b) $\int_0^1 g_3(t)dt = 0$, which eliminates the additive constant making g_3 identifiable with g_1. We use cubic splines to model g_1 and g_2 and periodic splines to model g_3. Specifically, let

$$W_2(per) = \{f : f^{(j)} \text{ are absolute continuous, } f^{(j)}(0) = f^{(j)}(1),\ j = 0,\ 1,\ \int_0^1 (f^{(2)}(t))^2 dt < \infty\}.$$

We assume that $\mathcal{H}_1 = W_2[0,1]$, $\mathcal{H}_2 = W_2[0,1] \ominus \{1\}$ and $\mathcal{H}_3 = W_2(per) \ominus \{1\}$, where $W_2[0,1]$ is defined in (4). Both side conditions are dealt with in a natural way by removing constant functions from the model spaces. Since it is known that the periodic pattern is close to a linear combination of $\sin 2\pi t$ and $\cos 2\pi t$, we model g_3 using a periodic L-spline with $\mathcal{L} = D^2 + (2\pi)^2$. Then it is easy to check that $\mathcal{H}_{10} = \text{span}\{1, t - 0.5\}$, $R_{11}(s,t) = k_2(s)k_2(t) - k_4(s-t)$, $\mathcal{H}_{20} = \text{span}\{t - 0.5\}$, $R_{21}(s,t) = R_{11}$, $\mathcal{H}_{30} = \text{span}\{\sin 2\pi t, \cos 2\pi t\}$, and $R_{31}(s,t) = \sum_{v=1}^{\infty} \frac{2}{(2\pi)^4(1-v^2)^2} \cos(2\pi v(s-t))$.

We use average thickness and zero as initial values for g_1 and g_2. To get initial values for g_3, we fit the centered data (thickness-mean(thickness)) with a periodic L-spline against csmonth. The fitted values are saved as f3.ini. Now we are ready to fit model (24) using nnr.

```
> nnr(thick~g1(csyear)+exp(g2(csyear))*g3(csmonth),
  func=list(g1(x)~list(~I(x-.5),cubic(x)),
            g2(x)~list(~I(x-.5)-1,cubic(x)),
            g3(x)~list(~sin(2*pi*x)+cos(2*pi*x)-1,lspline(x,type="sine0"))),
  data=ozone.dat, start=list(g1=mean(thick),g2=0,g3=f3.ini), spar=''m'')
```

The fit converged after 4 iterations with $\hat{\sigma} = 15.62$. Figure 1 shows the estimates of g_1 and g_2 and their approximate 95% Bayesian confidence intervals. These intervals indicate that g_2 is not significantly different from zero. Thus an additive model of month and year seems sufficient. g_1 can not be reduced to a linear function. We conclude that the mean ozone thickness changed over years and the amplitude remained the same. Figure 2 shows the estimate of g_3 and its decompositions into $\mathcal{H}_{30}$ (the parametric part) and $\mathcal{H}_{31}$ (the smooth part). Approximate 95%

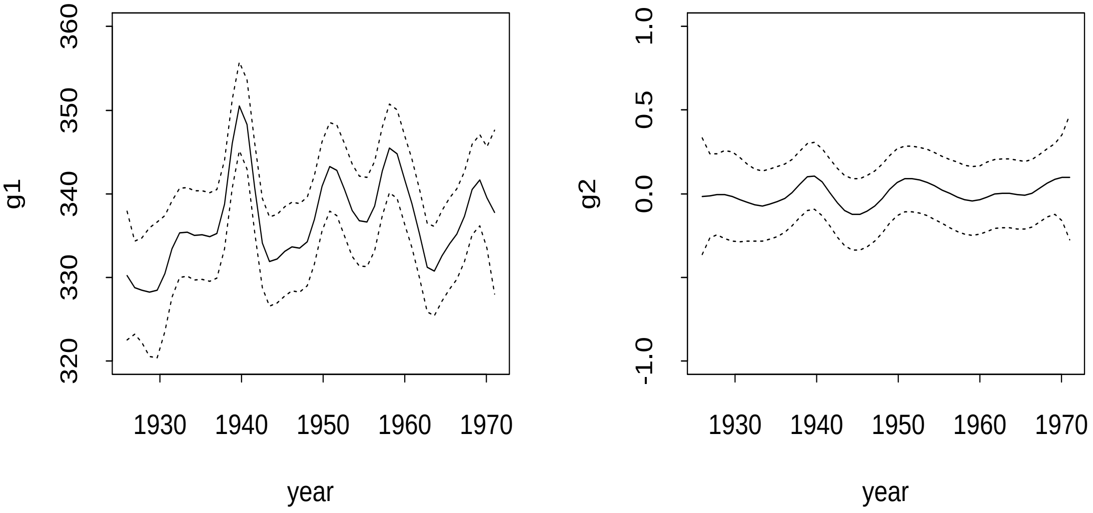

Figure 1: Plots of the estimates of g_1 (left) and g_2 (right) with their approximate 95% Bayesian confidence intervals (dotted lines).

Bayesian confidence intervals are constructed for each component. Since the smooth part is significantly different from zero, we conclude that a simple sinusoidal model is not sufficient to describe the periodic pattern.

8 Simulations

We conduct two sets of simulations to evaluate the performance of the estimation and inference methods developed in this article.

In the first set of simulations, data were generated from a simple nonlinear model

$$y_i = \exp(g(t_i)) + \epsilon_i, \quad i = 1, \cdots, n,$$

where $g(t) = \sin(2\pi t)$, $t_i = i/n$ and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We used a factorial design with two choices of n (50 and 100) and two choices of σ (.5 and 1). For each combination of n and σ, we ran the simulation until it reached 100 converged replicates. Both the iterative GCV and iterative GML were used to select smoothing parameters. To evaluate the performance, we calculated (a) bias of σ; (b) the mean square error (MSE) for σ; and (c) the integrated MSE for g: IMSE=$\int_0^1 (g(s) - \hat{g}(s))^2 ds$.

When the iterative GCV method was used, there were 11, 9, 4 and 3 nonconvergent runs for

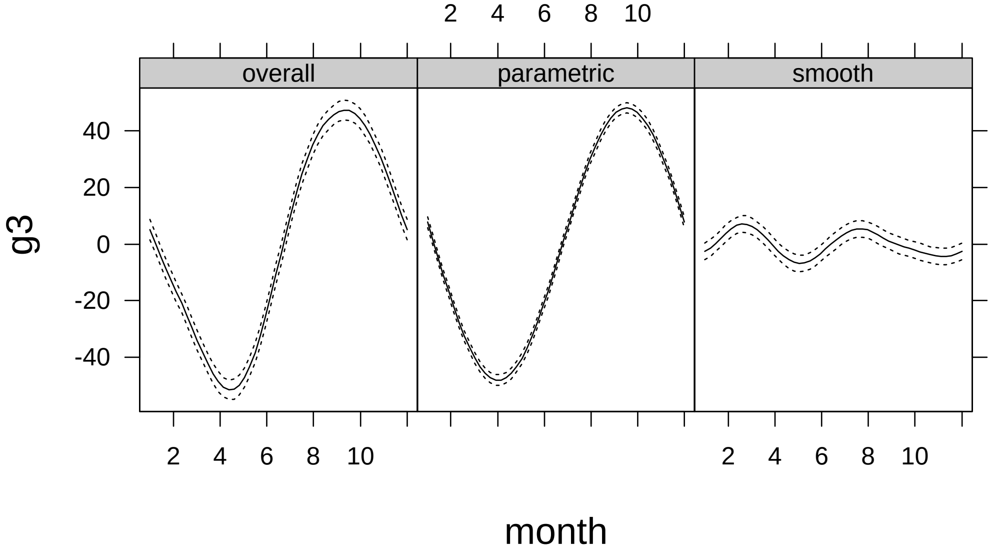

Figure 2: Plots of the estimate of g_3 (left) and its decompositions into $\mathcal{H}_{30}$ (middle) and $\mathcal{H}_{31}$ (right). The dotted lines are approximate 95% Bayesian confidence intervals.

the combinations of ($n = 50$, $\sigma = 1$), ($n = 50$, $\sigma = .5$), ($n = 100$, $\sigma = 1$) and ($n = 100$, $\sigma = .5$) respectively. When the iterative GML method was used, there were two non-convergent runs: one for ($n = 50$, $\sigma = 1$) and one for ($n = 50$, $\sigma = 0.5$). Thus the GML method was more stable. We calculated GCV scores (19) on a fine grid of $\log_{10}(n\lambda)$ for a typical nonconvergent case ($n = 100$, $\sigma = .5$). For any fixed λ, a_{ij} in (19) were evaluated at $\hat{g}$. The GCV scores are shown in Figure 3. We can see that the approximate GCV score has two local minima. λ chosen by the iterative GCV method jumped between -4.08 and -2.52, thus causing nonconvergence.

Table 1 summarizes the bias and MSE for $\hat{\sigma}$ and the IMSE for $\hat{g}$. Estimates are better for larger sample size and/or smaller variance. Overall, both GCV and GML methods lead to good estimates, and the GML method performs slightly better. Note that σ is consistently underestimated with very small bias.

Figure 4 shows the coverages of the 95% approximate Bayesian confidence intervals and bootstrap confidence intervals. The bootstrap confidence intervals were constructed based on the estimates of 100 bootstrap samples. Smoothing parameters were selected by the iterative GML method. Overall, the bootstrap confidence intervals are comparable to the approximate Bayesian confidence intervals, and are slightly better when the sample size is small. The performance obviously depends on the sample size and error variance. Variations are large especially when the sample size is small. Curvature is large for this simulation. Thus certain

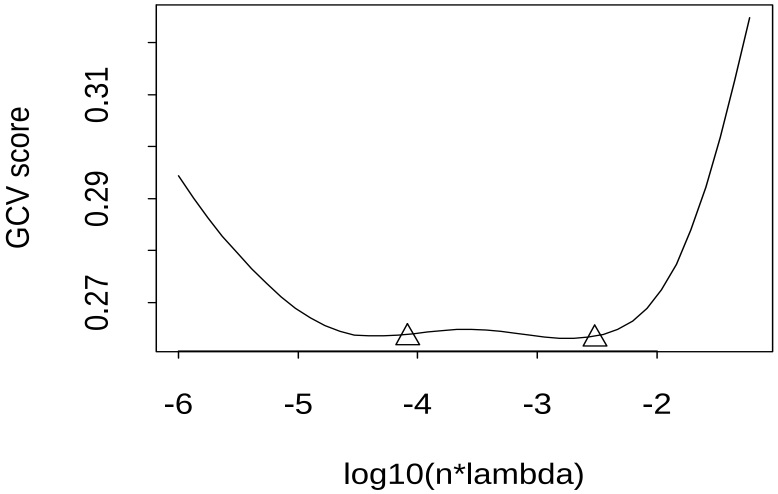

Figure 3: Approximate GCV curve for a typical non-convergent case. The iterative GCV method selects the smoothing parameter alternatively between two values marked by Δ.

adjustments may improve the performance. Figure 5 shows the pointwise bias, variance, MSE and coverages of the approximate 95% Bayesian confidence intervals for the simulation setting with $n = 100$ and $\sigma = 0.5$. It is clear that biases, variances and coverages depend on the curvatures of both $\mathcal{N}$ and g.

The second group of simulations imitate model (24) used for the ozone example. Data were generated from

$$y_{ij} = g_1(s_j) + \exp(g_2(s_j))g_3(t_i) + \epsilon_{ij}, \quad i, j = 1, \cdots, n,$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $t_i = i/n$ and $s_j = j/n$. Let $g_1(s) = s^4$, $g_2(s) = -\frac{1}{2}(s - 0.5)$ and $g_3(t) = \exp(\frac{1}{4})(\beta(t, 6, 4) - 1)$, where $\beta(t, a, b)$ is the density function of the Beta-distribution

Table 1: Bias and MSE of $\hat{\sigma}$, and IMSE of $\hat{g}$

		$\sigma = 0.5$			$\sigma = 1$		
λ	n	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)	IMSE($\hat{g}$)	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)	IMSE($\hat{g}$)
GCV	50	-0.0154	0.0041	0.0490	-0.0467	0.0162	0.2107
GML	50	-0.0061	0.0029	0.0373	-0.0385	0.0166	0.2959
GCV	100	-0.0033	0.0014	0.0264	-0.0127	0.0059	0.0866
GML	100	-0.0018	0.0012	0.0232	-0.0083	0.0054	0.0790

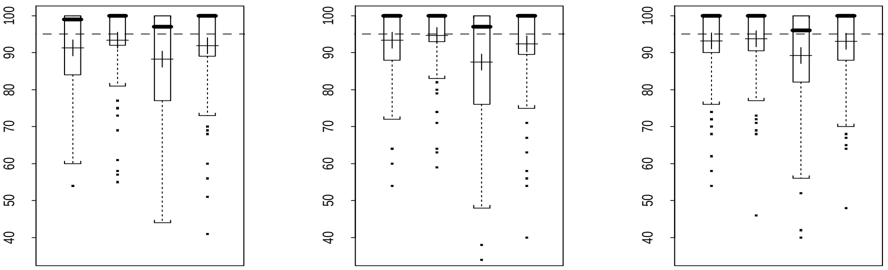

Figure 4: Boxplots of coverages of 95% approximate Bayesian confidence intervals based on the GCV (left) and GML (middle) methods and bootstrap confidence intervals (right). The boxes from left to right in each plot correspond to combinations of ($\sigma = 0.5$, $n = 50$), ($\sigma = 0.5$, $n = 100$), ($\sigma = 1$, $n = 50$) and ($\sigma = 1$, $n = 100$) respectively. Plusses and horizontal bars inside each box represent mean and median coverages respectively. The dashed line in each plot represents the nominal coverage.

with parameters a and b. Note that $\int_0^1 g_2(t)dt = 0$ and $\int_0^1 g_3(t)dt = 0$. Thus the identifiability conditions are satisfied. We used a factorial design with two choices of n (15 and 20) and two choices of σ (0.3 and 0.5). All 4 combinations were repeated 100 times. The iterative GML method was used to choose smoothing parameters. All simulations converged quickly after about 4 iterations.

Table 2 shows the bias and MSE for $\hat{\sigma}$ and the IMSE for $\hat{g}_1$, $\hat{g}_2$, and $\hat{g}_3$. We can see that all estimates are very close to the true parameters and functions. As expected, smaller error variance and/or larger sample size leads to better estimates.

Table 2: Bias and MSE of $\hat{\sigma}$, and IMSE of $\hat{g}_1$, $\hat{g}_2$ and $\hat{g}_3$.

σ	n	Bias($\hat{\sigma}$)	MSE($\hat{\sigma}$)	IMSE($\hat{g}_1$)	IMSE($\hat{g}_2$)	IMSE($\hat{g}_3$)
0.3	15	-0.0032	0.0002	0.0021	0.0005	0.0028
0.5	15	-0.0029	0.0006	0.0049	0.0012	0.0065
0.3	20	-0.0027	0.0001	0.0012	0.0002	0.0017
0.5	20	-0.0012	0.0004	0.0027	0.0006	0.0042

We plot in Figure 6 the true functions together with their estimates corresponding to the minimum, lower quartile, median, upper quartile, and maximum IMSE for the case $n = 15$ and $\sigma = 0.3$. We can see that these estimates are good even for the one with the largest IMSE.

Figure 7 summarizes coverage of the approximate confidence intervals. For linear spline models (1), it is known that Bayesian confidence intervals for a cubic spline are simultaneous

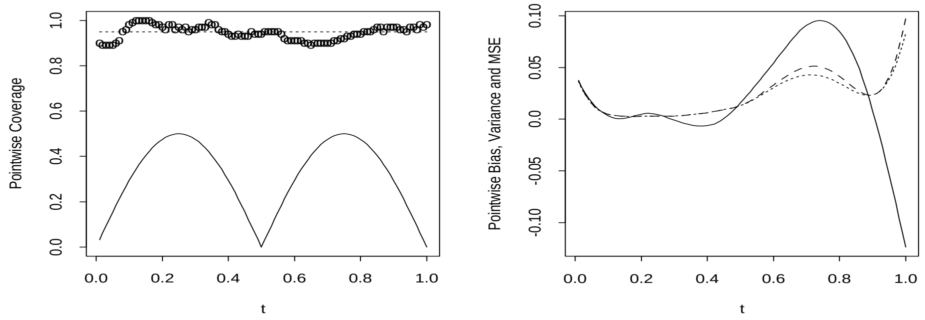

Figure 5: Left: pointwise coverages of 95% Bayesian confidence intervals (points) and $|g''|$ scaled into [0,.5] (solid line). Right: pointwise bias (solid line), variance (dotted line), and MSE (dashed line).

confidence intervals when the true function is in the null space. The middle plot in Figure 7 suggests that this result remains true for NNRMs.

9 Discussions

We have introduced NNRMs as extensions of the simple nonparametric regression and GAM-type models. They are also extensions of the parametric nonlinear regression models by allowing parameters in infinite dimensional spaces. Therefore these models can be used to test if a nonlinear regression model is appropriate. We briefly discussed the potential power by combining L-splines with nonlinear transformations in Example 4. Their applications for testing nonlinear regression models will be pursued in the future.

Throughout this paper we have assumed that the random errors are independent. It is not difficult to generalize NNRMs to the situation with correlated random errors. Specifically, we may assume that $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{W}^{-1})$ in (3). Then we can estimate unknown functions by penalized likelihood. This is to replace the least squares in (7) by weighted least squares. The result in Section 3.1 still holds, and the methods in Sections 3.2 and 3.3 can be modified similarly. The selection of smoothing parameters needs special attention when random errors are correlated. The GCV and GML methods for correlated data in Wang (1998b) can be used at each iteration.

Since the unknown nonparametric functions are in very flexible model spaces, it sometimes seems that the nonlinear function is redundant. For example, a multiplicative model $\mathcal{N}(g_1, g_2) = g_1 g_2$ can be transformed into an additive model by taking logarithm. Whenever applicable, this kind of transformations should be used. However, a simple transformation may not be feasible. For example, the logarithm may not be used when one of the functions is not positive.

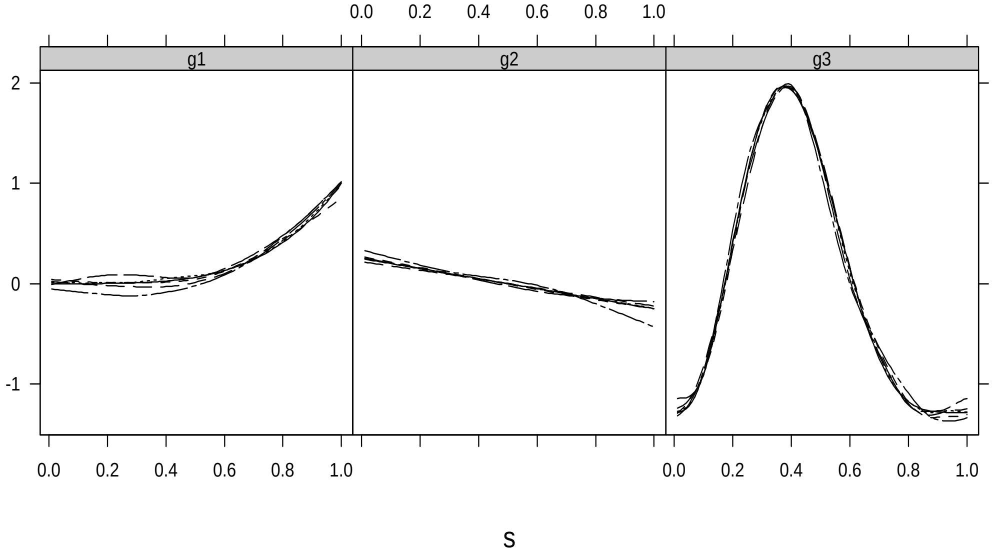

Figure 6: Estimate of g_1 (left), g_2 (middle) and g_3 (right) for the case $\sigma = 0.3$ and $n = 15$. Solid lines are the true functions. Dotted and dashed lines are estimates corresponding to minimum, lower quartile, median, upper quartile, and maximum of the IMSE.

Furthermore, it is important to understand that a transformation of the response variable also changes the assumption on the random errors.

Our methods for NNRMs have the same problems as those for nonlinear regression models: (a) convergence of iterative procedures and (b) accuracy of approximate inference based on linearization. Note that the Gauss-Newton method in this paper is acturally closer to the Levenberg-Marquardt method due to the penalty term. Thus this procedure is quite stable. According to our experience, nonconvergence is mainly caused by multiple local minima of a smoothing parameter selection criterion. Therefore more careful implementation of the iterative procedure is necessary. Large variations in coverages of the approximate Bayesian confidence intervals are caused by nonlinearity of the operator. Further research is required on the measure of curvature, its impact on the approximate inference, possible corrections and diagnostic tools.

APPENDIX A: Proof of Theorem 1

For simplicity of notation, we consider the case $r = 1$. The proof for $r > 1$ is similar. Let $\mathcal{S} = \mathcal{H}_0 \oplus \text{span}\{\xi_1, \cdots, \xi_n\}$, and $\mathcal{S}^c$ be the orthogonal complement of $\mathcal{S}$. For any $g \in \mathcal{H}$, let $g = g_1 + g_2$, where $g_1 \in \mathcal{S}$ and $g_2 \in \mathcal{S}^c$. Then

$$\mathcal{L}_i g =< \delta_i, g >=< \delta_i, g_1 > + < \delta_i, g_2 >=< \delta_i, g_1 > + < \xi_i, g_2 >=< \delta_i, g_1 >= \mathcal{L}_i g_1.$$

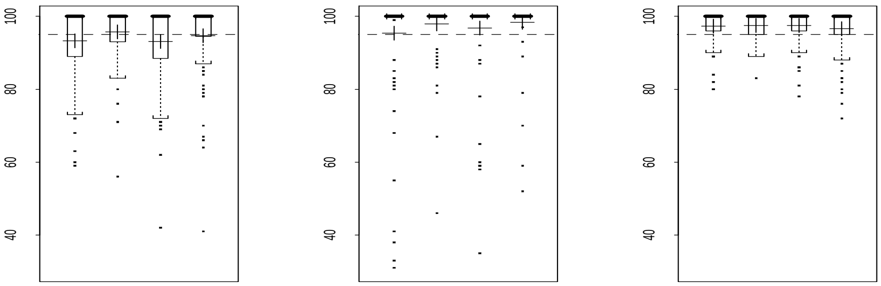

Figure 7: Boxplots of coverage of 95% approximate Bayesian confidence intervals for $\hat{g}_1$ (left), $\hat{g}_2$ (middle) and $\hat{g}_3$ (right). Four boxes from left to right in each plot correspond to the combinations of ($\sigma = 0.3$, $n = 15$), ($\sigma = 0.3$, $n = 20$), ($\sigma = 0.5$, $n = 15$) and ($\sigma = 0.5$, $n = 20$) respectively. Plusses and horizontal bars inside each box represent mean and median coverages respectively. The dashed line in each plot represents the nominal coverage.

Let $LS(\boldsymbol{y}; \eta(\mathcal{L}_1 g, \cdots, \mathcal{L}_n g))$ be the least squares. Then the penalized least squares

$$\begin{aligned}
& PLS(\boldsymbol{y}; \eta_1(\mathcal{L}_1 g), \cdots, \eta_n(\mathcal{L}_n g)) \\
= \ & LS(\boldsymbol{y}; \eta_1(\mathcal{L}_1 g), \cdots, \eta_n(\mathcal{L}_n g)) + n\lambda ||P_1 g||^2 \\
= \ & LS(\boldsymbol{y}; \eta_1(\mathcal{L}_1 g_1), \cdots, \eta_n(\mathcal{L}_n g_1)) + n\lambda (||P_1 g_1||^2 + ||P_1 g_2||^2) \\
\geq \ & LS(\boldsymbol{y}; \eta_1(\mathcal{L}_1 g_1), \cdots, \eta_n(\mathcal{L}_n g_1)) + n\lambda ||P_1 g_1||^2 \\
= \ & PLS(\boldsymbol{y}; \eta_1(\mathcal{L}_1 g_1), \cdots, \eta_n(\mathcal{L}_n g_1)).
\end{aligned}$$

Equality holds iff $||P_1 g_2|| = ||g_2|| = 0$. Thus the minimizers of (7) fall in $\mathcal{S}$, which can be represented by (9).

APPENDIX B: Gauss-Newton and Newton-Raphson Procedures

For simplicity of notation, we represent these procedures for $r = 1$. Extension to the case of multiple functions is straightforward. Let $\boldsymbol{c}_-$ and $\boldsymbol{d}_-$ be the current estimate of $\boldsymbol{c}$ and $\boldsymbol{d}$. Let $\hat{g}_-$ be the current estimate of g calculated from (9) with $\boldsymbol{c}$ and $\boldsymbol{d}$ replaced by $\boldsymbol{c}_-$ and $\boldsymbol{d}_-$, and $\boldsymbol{V} = diag(\eta'_1(\mathcal{L}_1 \hat{g}_-), \cdots, \eta'_n(\mathcal{L}_n \hat{g}_-))$.

Gauss-Newton algorithm: we expand $\eta_i(\mathcal{L}_i g)$ to the first order at $\hat{g}_-$

$$\eta_i(\mathcal{L}_i g) \approx \eta_i(\mathcal{L}_i \hat{g}_-) - \eta'(\mathcal{L}_i \hat{g}_-) \mathcal{L}_i \hat{g}_- + \eta'(\mathcal{L}_i \hat{g}_-) \mathcal{L}_i g. \tag{25}$$

Let $\check{y}_i = y_i - \eta_i(\mathcal{L}_i \hat{g}_-) + \eta'_i(\mathcal{L}_i \hat{g}_-) \mathcal{L}_i \hat{g}_-$ and $\check{\boldsymbol{y}} = (\check{y}_1, \cdots, \check{y}_n)^T$. Then (10) is approximated by

$$||\check{\boldsymbol{y}} - \boldsymbol{V}(\boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c})||^2 + n\lambda \boldsymbol{c}^T \boldsymbol{\Sigma} \boldsymbol{c}. \tag{26}$$

Let $\check{\boldsymbol{T}} = \boldsymbol{V}\boldsymbol{T}$, $\check{\boldsymbol{\Sigma}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}$, $\check{\boldsymbol{c}} = \boldsymbol{V}^{-1}\boldsymbol{c}$ and $\check{\boldsymbol{d}} = \boldsymbol{d}$. Then the Gauss-Newton method updates $\boldsymbol{c}$ and $\boldsymbol{d}$ by solving

$$\begin{aligned} (\check{\boldsymbol{\Sigma}} + n\lambda\boldsymbol{I})\check{\boldsymbol{c}} + \check{\boldsymbol{T}}\check{\boldsymbol{d}} &= \check{\boldsymbol{y}}, \\ \check{\boldsymbol{T}}^T\check{\boldsymbol{c}} &= 0. \end{aligned} \tag{27}$$

Equations (27) can be solved by `RKPACK` (Gu, 1989), and λ can be estimated by the GCV or GML method.

Newton-Raphson algorithm. Let $I(\boldsymbol{c},\boldsymbol{d}) = \sum_{i=1}^n (y_i - \eta_i(\sum_{u=1}^m d_u\mathcal{L}_i\phi_u + \sum_{v=1}^n c_v\mathcal{L}_i\xi_v))^2/2$, $\boldsymbol{\eta} = (\eta_1(\mathcal{L}_1\hat{g}_-),\cdots,\eta_n(\mathcal{L}_n\hat{g}_-))^T$, $\boldsymbol{u} = -\boldsymbol{V}(\boldsymbol{y}-\boldsymbol{\eta})$, $\boldsymbol{E} = diag(\eta_1''(\mathcal{L}_1\hat{g}_-),\cdots,\eta_n''(\mathcal{L}_n\hat{g}_-))$ and $\boldsymbol{\Lambda} = \boldsymbol{V}^2 - \boldsymbol{E}$. Then

$$\begin{aligned} \partial I/\partial\boldsymbol{c} &= -\boldsymbol{\Sigma}\boldsymbol{V}(\boldsymbol{y}-\boldsymbol{\eta}) = \boldsymbol{\Sigma}\boldsymbol{u}, \\ \partial I/\partial\boldsymbol{d} &= -\boldsymbol{T}^T\boldsymbol{V}(\boldsymbol{y}-\boldsymbol{\eta}) = \boldsymbol{T}^T\boldsymbol{u}, \\ \partial^2 I/\partial\boldsymbol{c}\partial\boldsymbol{c}^T &= \boldsymbol{\Sigma}\boldsymbol{V}^2\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{E}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma}, \\ \partial^2 I/\partial\boldsymbol{c}\partial\boldsymbol{d}^T &= \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{T}, \\ \partial^2 I/\partial\boldsymbol{d}\partial\boldsymbol{d}^T &= \boldsymbol{T}^T\boldsymbol{\Lambda}\boldsymbol{T}. \end{aligned}$$

The Newton-Raphson iteration satisfies the following equations

$$\begin{pmatrix} \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma} + n\lambda\boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{T} \\ \boldsymbol{T}^T\boldsymbol{\Lambda}\boldsymbol{\Sigma} & \boldsymbol{T}^T\boldsymbol{\Lambda}\boldsymbol{T} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} - \boldsymbol{c}_- \\ \boldsymbol{d} - \boldsymbol{d}_- \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\Sigma}\boldsymbol{u} - n\lambda\boldsymbol{\Sigma}\boldsymbol{c}_- \\ -\boldsymbol{T}^T\boldsymbol{u} \end{pmatrix}. \tag{28}$$

Similar to Gu (1990), $\boldsymbol{g} = \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c}$ is always unique as long as $\boldsymbol{T}$ is of full column rank, which we assume to be true. Thus we only need *a* solution of (28). If $\boldsymbol{\Sigma}$ is nonsingular, (28) is equivalent to

$$\begin{aligned} (\boldsymbol{\Lambda}\boldsymbol{\Sigma} + n\lambda\boldsymbol{I})\boldsymbol{c} + \boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{d} &= \boldsymbol{\Lambda}\hat{\boldsymbol{g}}_- - \boldsymbol{u}, \\ \boldsymbol{T}^T\boldsymbol{c} &= 0, \end{aligned} \tag{29}$$

where $\hat{\boldsymbol{g}}_- = (\mathcal{L}_1\hat{g}_-,\cdots,\mathcal{L}_n\hat{g}_-)^T$. If $\boldsymbol{\Sigma}$ is singular, any solution to (29) is also a solution to (28). Assume that $\boldsymbol{\Lambda}$ is positive definite. Let $\breve{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Lambda}^{1/2}$, $\breve{\boldsymbol{T}} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{T}$, $\breve{\boldsymbol{c}} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{c}$, $\breve{\boldsymbol{d}} = \boldsymbol{d}$ and $\breve{\boldsymbol{y}} = \boldsymbol{\Lambda}^{-1/2}(\boldsymbol{\Lambda}\hat{\boldsymbol{g}}_- - \boldsymbol{u})$. Then (29) is simplified to

$$\begin{aligned} (\breve{\boldsymbol{\Sigma}} + n\lambda\boldsymbol{I})\breve{\boldsymbol{c}} + \breve{\boldsymbol{T}}\breve{\boldsymbol{d}} &= \breve{\boldsymbol{y}}, \\ \breve{\boldsymbol{T}}^T\breve{\boldsymbol{c}} &= 0, \end{aligned}$$

which again can be solved by `RKPACK`. Note that when $\boldsymbol{E}$ is ignored, $\boldsymbol{\Lambda} = \boldsymbol{V}^2$. Then the Newton-Raphson method is the same as the Gauss-Newton method.

APPENDIX C: Proof of Theorem 2

As described in Appendix B, the CGN method updates $\boldsymbol{c}$ and $\boldsymbol{d}$ by solving (26). On the other hand, without using the fact that the solution to (7) is in a finite dimensional space, the EGN method approximates the nonlinear operator first by (11). Then the solution to (11) is in a finite dimensional space and has the form (14). Thus we need to show that given the same current estimates, representation (14) is the same as (9) with $r = 1$, and coefficients are updated by the same equation (26).

Since $\mathcal{L}_i$ is bounded, its Fréchet differential exists and $\mathcal{L}_i' = \mathcal{L}_i$ (Theorem 9.2.6 in Debnath and Mikusinski, 1999). By the chain rule (Theorem 9.2.5 in Debnath and Mikusinski, 1999), the Fréchet differential of $\mathcal{N}_i$ is $\mathcal{N}_i' = \eta_i'(\mathcal{L}_i g)\mathcal{L}_i$. Denote $\tilde{g}_-$ as the current estimate of the EGN method and assume that $\tilde{g}_- = \hat{g}_-$. Then $\mathcal{D}_i = \mathcal{N}_i'|_{\tilde{g}_-} = \eta_i'(\mathcal{L}_i\tilde{g}_-)\mathcal{L}_i$. At the next iteration, the EGN method approximate $\eta_i(\mathcal{L}_i g)$ by

$$\eta_i(\mathcal{L}_i g) = \mathcal{N}_i(g) \approx \mathcal{N}_i(\tilde{g}_-) + \mathcal{D}_i(g - \tilde{g}_-) = \eta_i(\mathcal{L}_i\tilde{g}_-) - \eta_i'(\mathcal{L}_i\tilde{g}_-)\mathcal{L}_i\tilde{g}_- + \eta_i'(\mathcal{L}_i\tilde{g}_-)\mathcal{L}_i g,$$

which is the same as (25). In (14), it is easy to check that $\tilde{\xi}_j = \eta_j'(\mathcal{L}_j\tilde{g}_-)\xi_j$, and $\tilde{\boldsymbol{c}}$ and $\tilde{\boldsymbol{d}}$ are solutions to

$$||\tilde{\boldsymbol{y}} - (\tilde{\boldsymbol{T}}\tilde{\boldsymbol{d}} + \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{c}})||^2 + n\lambda\tilde{\boldsymbol{c}}^T\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{c}}, \tag{30}$$

where $\tilde{y}_i = y_i - \eta_i(\mathcal{L}_i\tilde{g}_-) + \eta_i'(\mathcal{L}_i\tilde{g}_-)\mathcal{L}_i\tilde{g}_-$, $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)^T$, $\tilde{\boldsymbol{T}} = \{\mathcal{D}_i\phi_j\}_{i=1}^{n}{}_{j=1}^{m}$, and $\tilde{\boldsymbol{\Sigma}} = \{\mathcal{D}_i\mathcal{D}_j R_1\}_{i,j=1}^n$. It is easy to check that $\tilde{\boldsymbol{y}} = \check{\boldsymbol{y}}$, $\tilde{\boldsymbol{T}} = \boldsymbol{V}\boldsymbol{T}$ and $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T$. Let $\boldsymbol{d} = \tilde{\boldsymbol{d}}$ and $\boldsymbol{c} = \boldsymbol{V}\tilde{\boldsymbol{c}}$, then equations (26) and (30) are the same.

References

[1] Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems From Many Fields for the Student and Research Worker*. Springer: Berlin: New York.

[2] Aronszajn, N. (1950). The Theory of Reproducing Kernel. *Transactions of the American Mathematical Society* **68**, 337-404.

[3] Bates, D. M. and Watts, D. G. (1981). Parametric Transformations for Improved Approximate Confidence Regions in Nonlinear Least Squares. *The Annals of Statistics* **9**, 1152-1167.

[4] Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications.* Wiely: New York: UK.

[5] Debnath, L. and Mikusinski, P. (1999). *Introduction to Hilbert Spaces with Applications.* Academic Press: London.

[6] Engl, H. W., Hanke, M. and Neubauer, A. (1996). *Regularization of Inverse Problems.* Kluwer: Dordrecht.

[7] Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker: New York.

[8] Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall: London.

[9] Gu, C. (1989). RKPACK and Its Application: Fitting Smoothing Spline Models. *Proceedings of the Statistical Computing Section, American Statistical Association*, 42-51.

[10] Gu, C. (1990). Adaptive Spline Smoothing in Non-Gaussian Regression Models. *Journal of the American Statistical Association* **85**, 801-807.

[11] Gu, C. (1992). Cross-validating Non-Gaussian Data. *Journal of Computational Graphic Statistics* **1**, 169-179.

[12] Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

[13] Gu, C. and Wahba, G. (1993). Smoothing Spline ANOVA with Component-wise Bayesian Confidence Intervals. *Journal of Computational and Graphical Statistics* **55**, 353-368.

[14] Hamilton, D. C., Watts, D. G. and Bates, D. M. (1982). Accounting for Intrinsic Nonlinearity in Nonlinear Regression Parameter Inference Regions. *The Annals of Statistics* **10**, 386-393.

[15] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall, London.

[16] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society, Series B* **55**, 757-796.

[17] Heckman, N. and Ramsay, J. O. (2000). Penalized Regression with Model-based Penalties. *Canadian Journal of Statistics* **28**, 241-258.

[18] Jiang, J. (2000). A Nonlinear Gauss-Seidel Algorithm for Inference about GLMM. *Computational Statistics* **15**, 229-241.

[19] Lee, Y. J. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619-673.

[20] Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* **46**, 673-687.

[21] Lusternik, L.A. and Sobolev, V. J. (1974). *Elements of Functional Analysis*. Hindustan Publishing Co., New Delhi.

[22] Miller, K. M. (1991). *Introduction to Differential Equations*. Prentice Hall: London.

[23] O'Sullivan, F. (1986). A Statistical Perspective on Ill-Posed Inverse Problems (with Discussion). *Statistical Science* **4**, 502-527.

[24] O'Sullivan, F. (1990). Convergence Characteristics of Methods of Regularization Estimators for Nonlinear Operator Equations. *SIAM Journal on Numerical Analysis* **27**, 1635-1649.

[25] O'Sullivan, F. and Wahba, G. (1985). A Cross Validated Bayesian Retrieval Algorithm for Nonlinear Remote Sensing Experiments. *Journal of Computational Physics* **59**, 441-455.

[26] O'Sullivan, F. and Yandell, B. S. and Raynor, W. J., Jr (1986). Automatic Smoothing of Regression Functions in Generalized Linear Models. *Journal of the American Statistical Association* **81**, 96-103.

[27] Ramsay, J. O. (1998). Estimating Smooth Monotone Functions. *Journal of the Royal Statistical Society, Series B* **60**, 365-375.

[28] Sachdev, P. L (1991). *Nonlinear Ordinary Differential Equations and Their Applications*. Marcel Dekker: New York.

[29] Vardi, Y. and Lee, D. (1993). From Image Deblurring to Optimal Investments: Maximum Likelihood Solutions for Positive Linear Inverse Problems (with discussions). *Journal of the Royal Statistical Society, Series B* **55**, 569-612.

[30] Wahba, G. (1983). Bayesian Confidence Intervals for the Cross-validated Smoothing Spline. *Journal of the Royal Statistical Society, Series B* **45**, 133-150.

[31] Wahba, G. (1987). Three Topics in Ill Posed Inverse Problems. In *Inverse and Ill-Posed Problems*,(M. Engl and G. Groetsch, eds.), 37-51. Academic Press.

[32] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics, V59, Philadelphia.

[33] Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1994). Smoothing Spline for Exponential Families, with Application to Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics* **23**, 1865-1895.

[34] Wang, Y. (1998*a*). Mixed-effects Smoothing Spline ANOVA. *Journal of the Royal Statistical Society, Series B* **60**, 159-174.

[35] Wang, Y. (1998*b*). Smoothing Spline Models with Correlated Random Errors. *Journal of the American Statistical Association* **93**, 341-348.

[36] Wang, Y. and Ke, C. (2002). Assist: A suite of S-Plus Functions Implementing Spline Smoothing Techniques, Manual for the ASSIST package. Available at

`http://www.pstat.ucsb.edu/faculty/yuedong/research.`

[37] Wang, Y. and Wahba, G. (1995). Bootstrap Confidence Intervals for Smoothing Spline Estimates And Their Comparison to Bayesian Confidence Intervals. *Journal of Statistical Computation and Simulation* **51**, 263-279.

[38] Xiang, D. and Wahba, G. (1996). A Generalized Approximate Cross Validation for Smoothing Splines with Non-Gaussian Data. *Statistica Sinica* **6**, 675-692.

[39] Ye, J. (1998). On Measuring And Correcting the Effects of Data Mining And Model Selection. *Journal of the American Statistical Association* **93**, 120-131.