Scandinavian Journal of Statistics, Vol. 37: 26–46, 2010
doi: 10.1111/j.1467-9469.2009.00655.x
© 2009 Board of the Foundation of the Scandinavian Journal of Statistics. Published by Blackwell Publishing Ltd.

Polynomial Spline Estimation for a Generalized Additive Coefficient Model

LAN XUE

Department of Statistics, Oregon State University

HUA LIANG

Department of Biostatistics, University of Rochester

ABSTRACT. We study a semiparametric generalized additive coefficient model (GACM), in which linear predictors in the conventional generalized linear models are generalized to unknown functions depending on certain covariates, and approximate the non-parametric functions by using polynomial spline. The asymptotic expansion with optimal rates of convergence for the estimators of the non-parametric part is established. Semiparametric generalized likelihood ratio test is also proposed to check if a non-parametric coefficient can be simplified as a parametric one. A conditional bootstrap version is suggested to approximate the distribution of the test under the null hypothesis. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed methods. We further apply the proposed model and methods to a data set from a human visceral Leishmaniasis study conducted in Brazil from 1994 to 1997. Numerical results outperform the traditional generalized linear model and the proposed GACM is preferable.

Key words: conditional bootstrap, generalized additive models, knots, maximum likelihood estimation, optimal rate of convergence, spline approximation

1. Introduction

The most common model used in analysing the relationship between a discrete response variable and covariates is the generalized linear model (McCullagh & Nelder, 1989). With a given link function, it models the relationship between the dependent and explanatory variables through a linear function form. However, many data that arise in a variety of disciplines, such as, economics, political science, geography, and epidemiology, require more flexible forms than the usual linearity. Recently, many non-parametric and semiparametric models have been proposed to relax the strict linear assumption in the generalized linear models, such as, the generalized additive models (Hastie & Tibshirani, 1990; Härdle *et al.*, 2004), the generalized varying coefficient models (Cai *et al.*, 2000), the generalized partially linear models (Green & Silverman, 1994; Härdle *et al.*, 2000; Liang & Ren, 2005), and the generalized partially linear single index models (Carroll *et al.*, 1997).

In this paper, we propose a new semiparametric model, namely the generalized additive coefficient model (GACM), which is an extension of varying coefficient models (Hastie & Tibshirani, 1993). Similar to the generalized varying coefficient model, it allows the coefficients of the linear covariates to depend on certain covariates, called tuning variables. But it further imposes an additive function form on the coefficient functions to circumvent the so-called 'curse-of-dimensionality' problem when the dimension of the tuning variables is large. As seen in section 2, the proposed GACM is flexible enough to include the aforementioned non-parametric and semiparametric models as special cases.

A motivation of this study comes from an analysis of an epidemiological data set. It consists of the human visceral Leishmaniasis (HVL) case numbers in 117 health zones in Belo Horizonte, Brazil from 1994 to 1997. HVL is mainly a rural disease that has become prevalent in recent years in Brazilian urban areas. The first human case of HVL was recorded in March 1989 in Sabará, a municipality located in the Belo Horizonte metropolitan region. Afterwards, in spite of the undertaken control actions, the disease spread into the city from the northeast. The annual human cases recorded in the years 1994, 1995, and 1996 were 29, 46, and 45, respectively. A total of 40 cases were already reported only in the first semester of 1997. As argued in Assunção (2003), the small number of cases in each area produced very unstable rates, preventing a more focused public health action. One of the main interests of the study is to model the disease diffusion over time to better monitor the disease and allocate resources for disease control. A possible approach is to model the HVL case numbers using a traditional Poisson regression model with a polynomial time trend, see (8). But Belo Horizonte is a large Brazilian city with great social, econometric, and geological diversity. Also the disease first appeared in the northwest, then spread into the city afterwards. Thus, the dynamic of disease progress over time is different over the whole space. A traditional Poisson model with constant coefficients, such as (8), over the whole space may not be able to capture this spatially varying phenomena. But this can be incorporated into the GACM, by allowing the coefficients of linear covariates to vary smoothly with the location indexes (latitude and longitude), see (7). Our analysis in section 6 shows that the GACM outperforms the generalized linear model in terms of both estimation and prediction.

In the least squares setting, Xue & Yang (2006a,b) considered estimation of the additive coefficient model for Gaussian data using both kernel and polynomial spline methods. In contrast, this paper studies estimation, and also testing, of the model for non-Gaussian data through maximizing the likelihood with polynomial spline smoothing. The convergence results of the maximum likelihood estimators in this paper are similar to those for regression established by Xue & Yang (2006b). But as Huang (1998) pointed out, it is more technically challenging to establish the rate of convergence for the maximum likelihood estimators, as it cannot be viewed simply as an orthogonal projection because of its non-linear structure. Another contribution of this paper is to propose an efficient testing procedure for the coefficient functions by combining polynomial spline smoothing with conditional bootstrapping.

The use of polynomial spline smoothing in generalized non-parametric models has been investigated in various contexts. Stone (1986) first obtained the convergence rate of polynomial spline estimates for the generalized additive model. Stone (1994) and Huang (1998) focused on polynomial spline estimation of the generalized functional ANOVA model, while Huang *et al.* (2000) and Huang & Liu (2006) considered the functional ANOVA model and the single-index model in proportional hazards regression via maximum partial likelihood estimation, respectively. Polynomial spline smoothing is a global smoothing method, which approximates the unknown functions via polynomial splines characterized by only a finite number of coefficients. After the spline basis is chosen, the coefficients can be estimated by an efficient one-step procedure of maximizing the likelihood function. It is computationally cheaper than kernel-based methods, where the maximizing has to be conducted repeatedly at every local data point. Thus, the application of polynomial spline smoothing in the current context is particularly computationally efficient.

The paper is organized as follows. Section 2 introduces the GACM. Section 3 gives an efficient polynomial spline estimation method for the proposed model. Mean square (or L_2) convergence results are established for the estimators under mild assumptions. Section 4 discusses a testing procedure of the coefficient functions via a conditional bootstrap approach. Simulation studies and an application of the proposed methods in a real data example are included in sections 5 and 6, respectively. Technical lemmas and proofs are given in the Appendix.

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

2. The model

In our definition of the generalized regression models, we follow the notation in Stone (1986, 1994), and Huang (1998). The set-up involves an exponential family of distributions of the form

$$\exp(B(\eta)y - C(\eta))\rho(\mathrm{d}y),$$

where $B(\cdot)$ and $C(\cdot)$ are known functions with $C(\eta) = \log \int_{\mathcal{R}} \exp[B(\eta)y]\rho(\mathrm{d}y)$, and ρ is a non-zero measure defined on $\mathcal{R}$ that is not concentrated on a single point. Correspondingly, the mean of the distribution is $\mu = A(\eta) = C'(\eta)/B'(\eta)$, where $B'(\cdot)$ and $C'(\cdot)$ are the first-order derivatives of $B(\cdot)$ and $C(\cdot)$, respectively. Equivalently, $\eta = A^{-1}(\mu)$ with the function A^{-1} being the link function.

Consider a random vector $(Y, \mathbf{X}, \mathbf{T})$, in which Y is a real-valued response variable, and $(\mathbf{X}, \mathbf{T})$ are predictor variables with $\mathbf{X} = (X_1, \ldots, X_{d_2})^{\mathrm{T}}$ and $\mathbf{T} = (T_1, \ldots, T_{d_1})^{\mathrm{T}}$. The conditional distribution of Y given $(\mathbf{X}, \mathbf{T})$ is connected to the above exponential family distribution through the assumption that

$$E(Y \mid \mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}) = A(\eta(\mathbf{x}, \mathbf{t})), \quad \eta(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} \left\{ \alpha_{l0} + \sum_{s=1}^{d_2} \alpha_{ls}(x_s) \right\} t_l. \tag{1}$$

Equation (1) is called the GACM. To assure that the components in (1) are identifiable, we impose the restriction $E\{\alpha_{ls}(X_s)\} = 0$, for $l = 1, \ldots, d_1, s = 1, \ldots, d_2$. As in most work on non-parametric smoothing, estimation of the functions $\{\alpha_{ls}\}_{l=1,s=1}^{d_1,d_2}$ is conducted on a compact support. Without loss of generality, let the compact set be $\chi = [0, 1]^{d_2}$.

The proposed GACM in (1) is quite general. It is flexible enough to cover a variety of situations. For example, when $d_2 = 0$, or equivalently, there are no predictors of $\mathbf{X}$, (1) becomes the generalized linear model. When $d_2 = 1$, (1) becomes the generalized varying coefficient model (Cai *et al.*, 2000). When $T_1 = \cdots = T_{d_1} =$ constant, (1) becomes the generalized additive model (Hastie & Tibshirani, 1990; Härdle *et al.*, 2004).

Similar to Huang (1998), if the conditional distribution of Y given $\mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}$ follows the exponential family distribution described before with $\eta = \eta(\mathbf{x}, \mathbf{t})$, then assumption (1) is satisfied and the log-likelihood function is given by $l(h, \mathbf{X}, \mathbf{T}, Y) = B(h(\mathbf{X}, \mathbf{T}))Y - C(h(\mathbf{X}, \mathbf{T}))$, for any function h defined on $\chi \times \mathcal{R}^{d_1}$. If the conditional distribution of Y given $\mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}$ does not follow the exponential family distribution, we can think of $l(h, \mathbf{X}, \mathbf{T}, Y)$ as a pseudo-log-likelihood. For simplicity, we refer to both cases as the log-likelihood functions.

3. Polynomial spline estimation

We propose to estimate the non-parametric functions in model (1) using the polynomial spline smoothing method. It involves an approximation of the non-parametric functions $\{\alpha_{ls}(\cdot)\}_{l=1,s=1}^{d_1,d_2}$ using polynomial splines.

For each of the tuning variable directions, i.e. $s = 1, \ldots, d_2$, we introduce a knot sequence $k_{s,n}$ on $[0, 1]$, with N_n interior knots, $k_{s,n} = \{0 = v_{s,0} < v_{s,1} < \cdots < v_{s,N_n} < v_{s,N_n+1} = 1\}$. For any non-negative integer p, we denote $\varphi_s = \varphi^p([0, 1], k_{s,n})$, the space of functions, whose element

(i) is a polynomial of degree p (or less) on each of the intervals $[v_{s,i}, v_{s,i+1})$ for $i = 0, \ldots, N_n - 1$, and $[v_{s,N_n}, v_{s,N_n+1}]$ and

(ii) is $p - 1$ continuously differentiable on $[0, 1]$ if $p \geq 1$.

A function that satisfies (i) and (ii) is called a polynomial spline, a piecewise polynomial connected smoothly on the interior knots. For example, a polynomial spline with degree $p = 0$

is a piecewise constant function, and a polynomial spline with degree $p=1$ is a piecewise linear function and continuous on $[0, 1]$. The polynomial spline space φ_s is determined by the degree of the polynomial p and the knot sequence $k_{s,n}$. Let $h_s=h_{s,n}=\max_{i=0,\dots,N_n}|v_{s,i+1}-v_{s,i}|$, which is called the mash size of $k_{s,n}$ and can be understood as a smoothness parameter like the bandwidth in the local polynomial context. Define $h=\max_{s=1,\dots,d_2}h_s$, which measures the overall smoothness.

To consistently estimate functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ in (1), which are centred with $E(\alpha_{ls}(X_s))=0$, we introduce empirically centred polynomial splines,

$$\varphi_s^{0,n}=\left\{g_s: g_s\in\varphi_s, \frac{1}{n}\sum_{i=1}^n g_s(X_{is})=0\right\}.$$

The basis of $\varphi_s^{0,n}$ can be conveniently constructed. For example, we have used the empirically centred truncated power basis in the implementation, i.e.

$$\left\{B_{sj}=b_{sj}-\frac{1}{n}\sum_{i=1}^n b_{sj}(X_{is})\right\}_{j=1}^{J_n},$$

where $J_n=N_n+p$, and $\{b_{s1},\dots,b_{sJ_n}\}$ is the truncated power basis given as:

$$\{x_s,\dots,x_s^p,(x_s-v_{s,1})_+^p,\dots,(x_s-v_{s,N_s})_+^p\},$$

in which $(x)_+^p=(x_+)^p$. If the functions $\{\alpha_{ls}(x_s)\}_{l=1,s=1}^{d_1,d_2}$ are smooth enough, then one can approximate them by polynomial splines $\{g_{ls}(x_s)\in\varphi_s^{0,n}\}_{l=1,s=1}^{d_1,d_2}$. That is, for each $l=1,\dots,d_1$, $s=1,\dots,d_2$, one has

$$\alpha_{ls}(x_s)\approx g_{ls}(x_s)=\sum_{j=1}^{J_n}c_{ls,j}B_{sj}(x_s),$$

with a set of coefficients $\{c_{ls,j}\}_{j=1}^{J_n}$. Denote $l(\eta(\mathbf{X},\mathbf{T}),Y)=B[\eta(\mathbf{X},\mathbf{T})]Y-C[\eta(\mathbf{X},\mathbf{T})]$, and $l_n(\eta)=\frac{1}{n}\sum_{i=1}^n l(\eta(\mathbf{X}_i,\mathbf{T}_i),Y_i)$ for any $\eta(\mathbf{X},\mathbf{T})=\sum_{l=1}^{d_1}\{\alpha_{l0}+\sum_{s=1}^{d_2}\alpha_{ls}(X_s)\}T_l$. Then, the log-likelihood function $l_n(\eta)$ can be approximated by

$$\begin{aligned} l_n(\eta)\approx l_n(\boldsymbol{\alpha}_0,\mathbf{c})&=\frac{1}{n}\sum_{i=1}^n l\left(\sum_{l=1}^{d_1}\left\{\alpha_{l0}+\sum_{s=1}^{d_2}g_{ls}(X_{is})\right\}T_{il},Y_i\right)\\ &=\frac{1}{n}\sum_{i=1}^n l\left(\sum_{l=1}^{d_1}\left\{\alpha_{l0}+\sum_{s=1}^{d_2}\sum_{j=1}^{J_n}c_{ls,j}B_{sj}(X_{is})\right\}T_{il},Y_i\right), \end{aligned} \tag{2}$$

in which the coefficients $\boldsymbol{\alpha}_0=(\alpha_{10},\dots,\alpha_{d_10})^{\mathrm{T}}$ and $\mathbf{c}=\{c_{ls,j}\}_{l=1,s=1,j=1}^{d_1,d_2,J_n}$ can be solved by maximizing the log-likelihood function, i.e.

$$(\hat{\boldsymbol{\alpha}}_0,\hat{\mathbf{c}})=\arg\max_{\boldsymbol{\alpha}_0,\mathbf{c}} l_n(\boldsymbol{\alpha}_0,\mathbf{c}). \tag{3}$$

Then, the resulting estimator of the functions is given as:

$$\hat{\alpha}_{ls}(x_s)=\sum_{j=1}^{J_n}\hat{c}_{ls,j}B_{sj}(x_s),$$

for $l=1,\dots,d_1, s=1,\dots,d_2$. As a result,

$$\hat{\eta}(\mathbf{x},\mathbf{t})=\sum_{l=1}^{d_1}\left\{\hat{\alpha}_{l0}+\sum_{s=1}^{d_2}\hat{\alpha}_{ls}(x_s)\right\}t_l. \tag{4}$$

The maximization of (3) can be easily carried out using existing software for generalized linear models. Furthermore, only one maximum likelihood procedure is needed to estimate

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

all the components in the coefficient function, which is much more computationally efficient than the kernel-based method where one needs to perform the maximum likelihood estimation at each local point. On the other hand, the next theorem shows that the polynomial spline estimators enjoy the same optimal rate of convergence as the kernel estimators. In the following, $\|\cdot\|$ denotes the theoretical L_2-norm defined by (9) in the Appendix.

Theorem 1

Under the assumptions (C1)–(C8) *in the Appendix, one has*

$$\|\hat{\eta}-\eta\| = O_p(h^{p+1}+\sqrt{1/nh}),$$

and for $l=1,\ldots,d_1, s=1,\ldots,d_2$,

$$|\hat{\alpha}_{l0}-\alpha_{l0}| = O_p(h^{p+1}+\sqrt{1/nh}), \quad \|\hat{\alpha}_{ls}-\alpha_{ls}\| = O_p(h^{p+1}+\sqrt{1/nh}).$$

Theorem 1 shows the mean square (or L_2) consistency of polynomial spline estimators. When the smoothing parameter h takes the optimal order of $n^{-1/(2p+3)}$, $\|\hat{\alpha}_{ls}-\alpha_{ls}\|_2 = O_p(n^{-(p+1)/(2p+3)})$, which is the optimal rate of convergence for univariate non-parametric functions. As a result, each of the d_2 dimensional coefficient functions $\alpha_l(\mathbf{x}) = \alpha_{l0} + \Sigma_{s=1}^{d_2}\alpha_{ls}(x_s)$ is also estimated at the univariate optimal rate. Therefore, the proposed GACM effectively avoids the 'curse-of-dimensionality' by assuming an additive structure on the coefficient functions.

3.1. Knot number selection

The proposed polynomial spline estimation procedure crucially depends on the appropriate choice of knot sequence $\{k_{s,n}\}_{s=1}^{d_2}$, and in particular, on the number of interior knots N_n. Here, we propose to select N_n using an Akaike information criterion (AIC) procedure. For knot location, we use either equally spaced knots or quantile knots (sample quantiles with the same number of observations between any two adjacent knots). A similar procedure was also used in Huang *et al.* (2002) and Xue & Yang (2006b).

According to theorem 1, the optimal order of N_n is $n^{1/(2p+3)}$. Thus, we propose to choose the optimal knot number, N_n^{opt}, from a neighbourhood of $n^{1/(2p+3)}$. For our examples in sections 5 and 6, we have used $[0.5N_r, \min(5N_r, Tb)]$, where $N_r = \text{ceiling}(n^{1/(2p+3)})$, and $Tb = \{n/(4d_1)-1\}/d_2$ to ensure that the total number of parameters in (2) is less than $n/4$. Let $\hat{\eta}_{N_n}(\cdot)$ be the estimator of $\eta(\cdot)$ with the number of knots N_n, and the resulting log-likelihood function $l_n(N_n) = \frac{1}{n}\sum_{i=1}^n l(\hat{\eta}_{N_n}(\mathbf{X}_i, \mathbf{T}_i), Y_i)$. Let $q_n = (1+d_2J_n)d_1$ be the total number of parameters in (2). Then the optimal knot number, N_n^{opt}, is the one that minimizes the AIC value. That is

$$N_n^{\text{opt}} = \underset{N_n\in[0.5N_r,\, \min(5N_r,\, Tb)]}{\text{argmin}} \{-2l_n(N_n)+2q_n\}.$$

4. Hypothesis testing

After fitting GACM (1), a natural question that arises is whether the coefficient functions $\{\alpha_{ls}\}_{l=1,s=1}^{d_1,d_2}$ are actually varying, or more generally, whether certain parametric models, such as polynomials, fit the non-parametric components. This leads us to consider hypothesis testing problems such as:

$$H_0: \alpha_{ls}(x_s) = \alpha_{ls}(x_{ls}, \boldsymbol{\theta}) \text{ versus } H_1: \alpha_{ls}(x_s) \neq \alpha_{ls}(x_{ls}, \boldsymbol{\theta}),$$

where $\alpha_{ls}(x_s, \boldsymbol{\theta}) = \sum_{k=0}^q \theta_k x_s^k$, and $\boldsymbol{\theta}$ is a vector of unknown parameters in the polynomial function. It includes testing whether the component α_{ls} is varying, in which $\alpha_{ls}(x_s, \boldsymbol{\theta}) = 0$. One

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

option is the non-parametric likelihood ratio test statistic (Fan *et al.*, 2001), which is defined as:

$$T_n = 2\{l_n(H_1) - l_n(H_0)\}, \tag{5}$$

in which $l_n(H_0)$ and $l_n(H_1)$ are the log-likelihood functions calculated under the null and alternative hypotheses, respectively. To be more specific, under the null hypothesis, we model α_{ls} as a polynomial of degree q and approximate all the other functions in the model with polynomial splines of degree p with $p \geq q$. Under the alternative hypothesis, all functions in the model are approximated with polynomial splines of degree p. We have used the AIC procedure in subsection 3.1 to choose the optimal knot number for the full GACM under the alternative hypothesis. Then the same number of knots is used for estimation of the non-parametric functions in the null model.

Theorem 2

Under the assumptions (C1)–(C8) *in the Appendix, one has, under* $H_0, T_n \to 0$ *in probability as* $n \to \infty$*; otherwise, there exists* $\delta > 0$*, such that* $T_n > \delta$ *with probability tending to one.*

The result of theorem 2 suggests rejecting H_0 for large T_n. To obtain an appropriate critical value, we approximate the null distribution of T_n using the conditional bootstrap method; see also Cai *et al.* (2000) and Fan & Huang (2005). Let $\{\hat{\alpha}^0_{l0}\}_{l=1}^{d_1}$ and $\{\hat{\alpha}^0_{ls}\}_{l=1,s=1}^{d_1,d_2}$ be the estimators of the constants and coefficient functions under H_0. Let the resulting estimator of $\eta(\mathbf{x},\mathbf{t})$ be $\hat{\eta}^0(\mathbf{x},\mathbf{t}) = \sum_{l=1}^{d_1}\{\hat{\alpha}^0_{l0} + \sum_{s=1}^{d_2}\hat{\alpha}^0_{ls}(x_s)\}t_l$. In the conditional bootstrap procedure, a total of B bootstrap samples are generated. In our examples given in sections 5 and 6, we have used $B = 500$. In each of the samples ($b = 1, \ldots, B$), the values of the independent variables $(\mathbf{X}_i, \mathbf{T}_i)$ are kept the same as the observed ones, while a bootstrap sample Y_i^b is generated from the distribution of Y, with $\eta(\mathbf{x},\mathbf{t})$ being $\hat{\eta}^0(\mathbf{X}_i,\mathbf{T}_i)$. Then, a test statistic T_n^b is computed from the bootstrap sample $(\mathbf{X}_i, \mathbf{T}_i, Y_i^b)_{i=1}^n$ using (5). In the implementation, the AIC knot selection procedure for the alternative model is performed for each bootstrap sample. The distribution of $\{T_n^b\}_{b=1}^B$ is used to approximate the distribution of the test statistic T_n. In particular, for a given level of significance α, the $(1-\alpha)100\%$ percentile of $\{T_n^b\}_{b=1}^B$ is used as the critical value.

5. Simulation study

In this section, we investigate the finite-sample performance of the proposed estimation and testing methods through two simulation studies. We use the averaged integrated squared error (AISE) to evaluate the performance of the function estimators $\{\hat{\alpha}_{ls}(\cdot)\}_{l,s=1}^{d_1,d_2}$, which is defined as:

$$\mathrm{AISE}(\hat{\alpha}_{ls}) = \frac{1}{n_{\mathrm{rep}}}\sum_{r=1}^{n_{\mathrm{rep}}}\mathrm{ISE}(\hat{\alpha}^r_{ls}), \quad \mathrm{ISE}(\hat{\alpha}^r_{ls}) = \frac{1}{n_{\mathrm{grid}}}\sum_{m=1}^{n_{\mathrm{grid}}}\{\alpha_{ls}(x_{ms}) - \hat{\alpha}^r_{ls}(x_{ms})\}^2,$$

where $\hat{\alpha}^r_{ls}(\cdot)$ denotes the estimator of $\alpha_{ls}(\cdot)$ in the rth replication for $r = 1, \ldots, n_{\mathrm{rep}}$, and $\{(x_{m1}, \ldots, x_{md_2})^{\mathrm{T}}\}_{m=1}^{n_{\mathrm{grid}}}$ are the grid points where the non-parametric functions are evaluated. In both examples, we have used the sample size $n = 250, 500, 750$, and the number of replications $n_{\mathrm{rep}} = 1000$. The non-parametric functions $\alpha_{ls}(\cdot)$ are all evaluated on a grid of equally spaced points $x_{ms}, m = 1, \ldots, n_{\mathrm{grid}}$ with $x_{1s} = 0.025, x_{n_{\mathrm{grid}},s} = 0.975$, and $n_{\mathrm{grid}} = 96$.

5.1. Logistic regression

Data sets are generated from a logistic regression model where the binary response variable Y_i has the distribution

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

$$\operatorname{logit} P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i, \mathbf{T}_i = \mathbf{t}_i) = \sum_{l=1}^{2} \left\{ \alpha_{l0} + \sum_{s=1}^{2} \alpha_{ls}(x_s) \right\} t_l,$$

in which $\alpha_{10} = 1, \alpha_{20} = 0, \alpha_{11}(x) = \alpha_{21}(x) = \sin(2\pi x), \alpha_{12}(x) = 0,$ and $\alpha_{22}(x) = 2x - 1$. The covariates $\mathbf{T}_i = (T_{i1}, T_{i2})^{\mathrm{T}}$ and $\mathbf{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$ are independently generated from the standard bivariate normal and uniform($[0, 1]^2$) distributions, respectively.

We have applied the proposed polynomial spline estimation method with both linear ($p = 1$) and cubic splines ($p = 3$). Estimation with other degrees such as quadratic spline ($p = 2$) can also be used, but give similar findings. We have used the AIC procedure to select the number of knots that are evenly placed over the ranges of x_{is}, for each $s = 1, \ldots, d_2$. Table 1 summarizes the means and standard errors (in the parentheses) of $\{\hat{\alpha}_{l0}\}_{l=1,2}$, and the AISE of $\{\hat{\alpha}_{ls}(\cdot)\}_{l=1,2}^{s=1,2}$ from both linear and cubic spline estimations. It shows that the two spline fits are generally comparable with cubic spline slightly better than the linear spline in smaller sample sizes. In both cases, the standard errors of the constant estimators and the AISE of the function estimators decrease as the sample size n increases, which confirms theorem 1. The typical estimated curves (whose ISE is the median in the 1000 replications) from linear polynomial spline estimation are plotted in Fig. 1, together with the pointwise 95 per cent confidence intervals when $n = 500$. It clearly shows that the linear spline fits are reasonably good.

Next we examine the proposed testing procedure and consider the following hypothesis:

$$H_0 : \alpha_{12}(\cdot) = 0 \ \ \textit{versus} \ \ H_1 : \alpha_{12}(\cdot) \neq 0. \tag{6}$$

The power of the test is evaluated under a sequence of alternative models, $H_1 : \alpha_{12}(x) = \lambda \sin(2\pi x)$, where λ controls the degree of departure from the null hypothesis, with $\lambda = 0$ corresponding to H_0. The value λ is taken to be a grid of equally spaced points on $[0, 1.5]$. Based on 1000 replications for the sample sizes $n = 250, 500$, and 750, Fig. 2 plots the power functions with significance level $\alpha = 0.05$. It shows that the power increases to 1 rapidly as λ increases. The powers at $\lambda = 0$ are $0.054, 0.056, 0.051$ for $n = 250, 500$, and 750, respectively, which are all close to the corresponding significance level.

5.2. *Poisson regression*

In this example, we consider a Poisson regression model with

$$\log E(Y_i \mid \mathbf{X}_i = \mathbf{x}_i, \mathbf{T}_i = \mathbf{t}_i) = \sum_{l=1}^{2} \left\{ \alpha_{l0} + \sum_{s=1}^{2} \alpha_{ls}(x_s) \right\} t_l,$$

where different forms of coefficient functions are considered, with $\alpha_{10} = 1, \alpha_{20} = 0$, and $\alpha_{11}(x) = 4x(1 - x) - 2/3, \alpha_{12}(x) = 0, \alpha_{21}(x) = \sin^2(\pi x) - 0.5, \alpha_{22}(x) = \mathrm{e}^{2x-1}/(\mathrm{e} - \mathrm{e}^{-1}) - 1/2$. The covariates are generated in the same way as in the logistic regression example.

Table 1. *Logistic regression example: the means and standard errors* (*in parentheses*) *of* $\hat{\alpha}_{10}$, *and* $\hat{\alpha}_{20}$ *and the averaged integrated squared errors of* $\hat{\alpha}_{11}(\cdot), \hat{\alpha}_{12}(\cdot), \hat{\alpha}_{21}(\cdot)$, *and* $\hat{\alpha}_{22}(\cdot)$ *from* 1000 *replications*

Methods	n	$\alpha_{10} = 1$	$\alpha_{20} = 0$	α_{11}	α_{12}	α_{21}	α_{22}
Linear spline	250	1.169 (0.321)	−0.009 (0.251)	0.4234	0.2182	0.3256	0.2684
	500	1.068 (0.157)	0.005 (0.134)	0.0953	0.0726	0.0804	0.0686
	750	1.047 (0.126)	0.003 (0.108)	0.0679	0.0451	0.0546	0.0418
Cubic spline	250	1.197 (0.304)	0.002 (0.236)	0.3531	0.2165	0.2429	0.2689
	500	1.082 (0.159)	0.007 (0.137)	0.0954	0.0799	0.0751	0.0682
	750	1.066 (0.125)	0.001 (0.111)	0.0688	0.0495	0.0527	0.0467

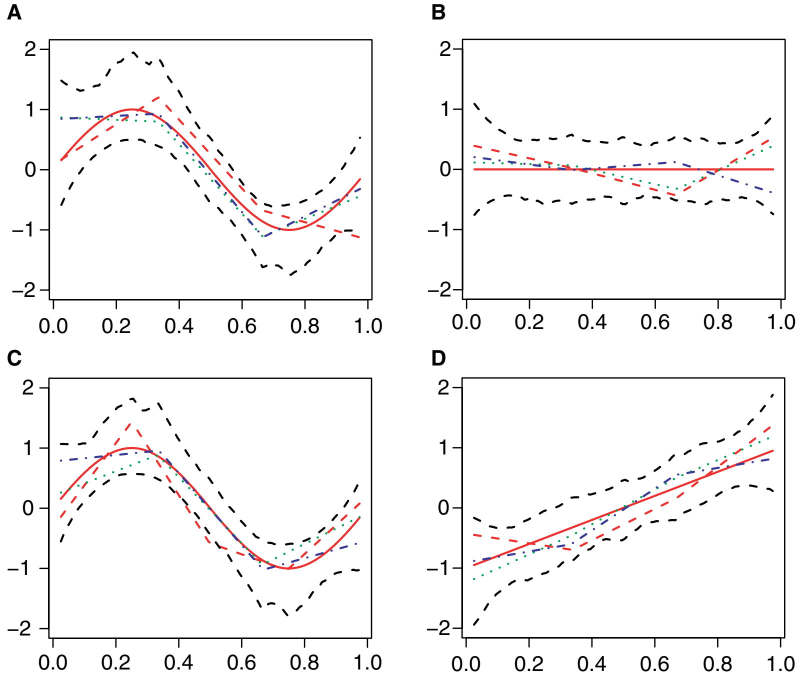

Fig. 1. Logistic regression example: plots of the typical estimated coefficient functions using the linear polynomial spline method: (A) α_{11}; (B) α_{12}; (C) α_{21}; (D) α_{22}. In each plot, the solid curve represents the true curve, the dashed curve is the typical estimated curve with $n=250$, the dotted curve is with $n=500$, and the dot-dash curve is with $n=750$. The two long-dashed curves are the pointwise 95 per cent confidence intervals when $n=500$.

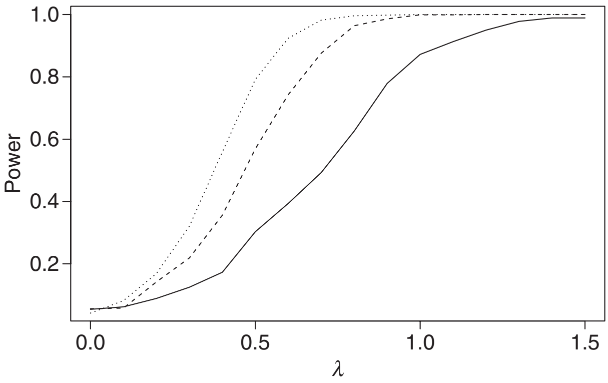

Fig. 2. Logistic example: the power function of the test statistics T_n is plotted against λ, for sample sizes $n=250$ (solid curve), 500 (dashed curve), and 750 (dotted curve). The significance level is 0.05.

Similar to that in the logistic regression example, we have used both linear spline ($p=1$) and cubic spline ($p=3$) estimations of the coefficient functions. Equally spaced knots are used with the number of interior knots chosen using the AIC procedure. The simulation results are summarized in Table 2, which contains the means and standard errors (in the parentheses) of $\{\hat{\alpha}_{l0}\}_{l=1,2}$, and the AISE of $\{\hat{\alpha}_{ls}(\cdot)\}_{l=1,2}^{s=1,2}$ from two spline fits. Similar to that in the logistic

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

Table 2. *Poisson regression example: the means and standard errors (in parentheses) of* $\hat{\alpha}_{10}$, *and* $\hat{\alpha}_{20}$ *and the averaged integrated squared errors of* $\hat{\alpha}_{11}(\cdot), \hat{\alpha}_{12}(\cdot), \hat{\alpha}_{21}(\cdot)$, *and* $\hat{\alpha}_{22}(\cdot)$ *using linear and cubic spline estimations from* 1000 *replications*

Methods	n	$\alpha_{10}=1$	$\alpha_{20}=0$	α_{11}	α_{12}	α_{21}	α_{22}
Linear spline	250	0.9923 (0.054)	0.0013 (0.061)	0.0099	0.0084	0.0168	0.0119
	500	0.9963 (0.033)	0.0022 (0.041)	0.0047	0.0036	0.0084	0.0054
	750	0.9957 (0.027)	0.0005 (0.033)	0.0030	0.0022	0.0056	0.0034
Cubic spline	250	0.9896 (0.053)	0.0016 (0.060)	0.0092	0.0091	0.0134	0.0112
	500	0.9945 (0.033)	0.0017 (0.039)	0.0041	0.0032	0.0053	0.0048
	750	0.9941 (0.027)	0.0009 (0.032)	0.0031	0.0024	0.0042	0.0037

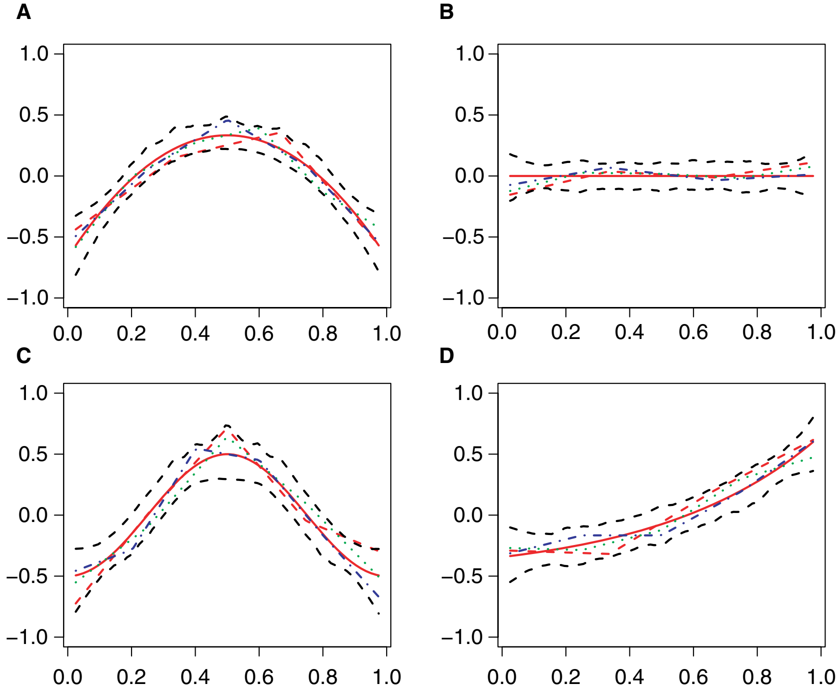

Fig. 3. Poisson regression example: plots of the typical estimated coefficient functions using the linear polynomial spline method: (A) α_{11}; (B) α_{12}; (C) α_{21}; (D) α_{22}. In each plot, the solid curve represents the true curve, the dashed curve is the typical estimated curve with $n=250$, the dotted curve is with $n=500$, and the dot-dash curve is with $n=750$. The two long-dashed curves are the pointwise 95 per cent confidence intervals when $n=500$.

regression example, Table 2 shows the convergence of both $\{\hat{\alpha}_{l0}\}_{l=1,2}$ and $\{\hat{\alpha}_{ls}(\cdot)\}_{l=1,2}^{s=1,2}$, as n increases. It again collaborates theorem 1. The typical estimated curves from the linear spline method with their pointwise 95 per cent confidence intervals when $n=500$ in Fig. 3 show that the proposed spline method gives reasonable estimators of the coefficient functions. We also studied the performance of the proposed testing procedure for this Poisson regression. The same hypothesis (6) as in the logistic regression example is considered. The power is evaluated under the alternative models, $H_1: \alpha_{12}(x)=\lambda[4x(1-x)-2/3]$, with λ being a grid of equally spaced points on $[0, 1.5]$. Based on 1000 replications for the sample size $n=250, 500$, and 750, Fig. 4 plots the power functions with significance level $\alpha=0.05$. The powers at $\lambda=0$ are $0.045, 0.053, 0.054$, respectively, which are close to the significance level.

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

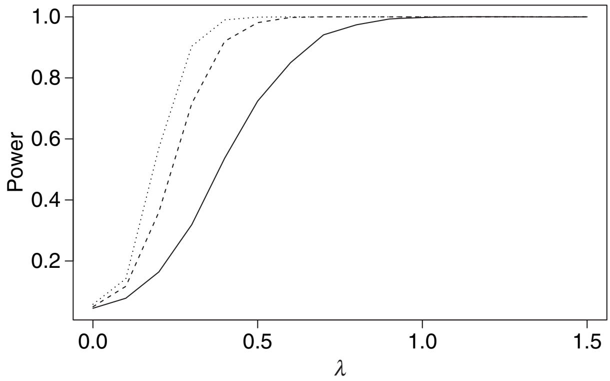

Fig. 4. Poisson regression example: the power function of the test statistics T_n is plotted against λ, for sample sizes $n=250$ (solid curve), 500 (dashed curve), and 750 (dotted curve). The significance level is 0.05.

6. Real data analysis

In this section, we apply the proposed GACM to analyse the data set from the HVL study introduced in section 1. The data consist of annual number of human HVL cases and total population counts for each of the 117 zones and each of the periods 1994/1995, 1995/1996, and 1996/1997. A period comprises the second semester of a year (starting 1 July) and the first semester of the following year (ending 30 June). For more information on the data, see Assunção *et al.* (2001). Belo Horizonte is a large Brazilian city with more than 2 million inhabitants. The spatial impacts are not necessarily homogenous over the whole area. Assunção *et al.* (2001) and Assunção (2003) took into account the varying spatial effect and used the Bayesian spatial varying parameter model to study the diffusion of the disease. Motivated by their analysis, we model the varying spatial effect by using the GACM as follows.

Let y_{it} be the annual counts of cases and P_{it} the risk population in each zone $(i), i=1,\ldots,117$, for 3 years $(t), t=0,1,2$. Time indexes $t=0,1,2$ represent the periods 1994/1995, 1995/1996, and 1996/1997, respectively. Similar to that in Assunção *et al.* (2001), and Assunção (2003), we assume, conditional on the relative risk $\exp(\lambda_{it})$, that the counts are independently distributed according to a Poisson distribution with mean $P_{it}\exp(\lambda_{it})$. A second-degree polynomial is assumed on λ_{it} to model the time trend. To allow for varying spatial effects, we further allow the coefficients of the polynomial terms to vary with the spatial coordinates of each zone (x_{i1}, x_{i2}). To be more specific, we assume

$$\lambda_{it}=\sum_{j=0}^{2}[\alpha_{j0}+\alpha_{j1}(x_{i1})+\alpha_{j2}(x_{i2})]t^{j}. \tag{7}$$

Model (7) allows the time profile to vary smoothly over space, thus effectively modelling the space–time interaction. The coefficient functions in (7) are estimated using both linear spline $(p=1)$ and quadratic spline $(p=2)$ with knot numbers selected by the AIC procedure as in subsection 3.1. Figure 5 plots the estimated coefficient functions. It shows that the two spline fits are very close. Therefore, for simplicity, we only report the results using linear spline in what follows. For comparison, we also consider a standard Poisson regression model with constant coefficients, i.e.

$$\lambda_{it}=c_0+c_1t+c_2t^2. \tag{8}$$

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

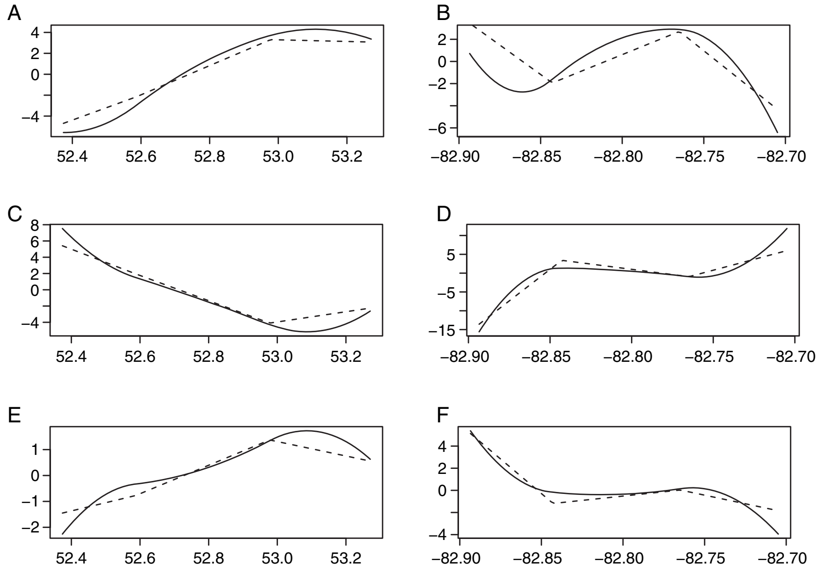

Fig. 5. The estimated coefficient functions in (7) using linear splines (dashed curve) and quadratic splines (solid curve). Plotted are (A) $\hat{\alpha}_{01}$, (B) $\hat{\alpha}_{02}$, (C) $\hat{\alpha}_{11}$, (D) $\hat{\alpha}_{12}$, (E) $\hat{\alpha}_{21}$, (F) $\hat{\alpha}_{22}$.

Fits are measured by their AIC, which is minus twice the maximized log-likelihood plus twice the number of parameters. Models (7) and (8) give AIC values 470.98 and 626.88 respectively, which indicate that (7) gives a better fit even with model complexity taken into account. Figure 6 graphically compares the residuals from two models, where the residuals are defined as $R_{it}=(y_{it}-\hat{y}_{it})/\sqrt{\hat{y}_{it}}$ with y_{it} and $\hat{y}_{it}$ being the observed and estimated annual count of cases for ith zone and tth year. We also compare the models by their prediction performances. We randomly select 15 zones from the 117 health zones. The observations taken during the last time period 1996/1997 from the selected 15 zones are left out for prediction, while the remaining observations are used for estimation. Then the averaged squared prediction errors (ASPE) are calculated. We replicated the prediction procedure ten times. Then the averaged ASPE from ten replications using (7) and (8) are reported, which are 1.14 and 1.51, respectively. That is, by efficiently taking the varying spatial effect into account, (7) not only provides better estimation performance, but also improves the prediction accuracy compared with the traditional Poisson regression model (8). Figure 7 plots the estimated HLV rates (per 100 thousands) from (7) in the health zones for each of the three periods.

Finally, one may ask whether the coefficient functions $\{\alpha_{ls}\}_{l=0,s=1}^{2,2}$ in (7) are all significantly different from zero, or whether (7) can be simplified with some of the coefficient functions deleted. For each $l=0,1,2$ and $s=1,2$, we apply the idea in section 4 to test the hypothesis: $H_0^{ls}:\alpha_{ls}(x_s)=0$. Based on 1000 bootstrap samples for each hypothesis, all coefficient functions are significantly different from 0 at level 0.05 with the p-values given as <0.0001, <0.0001, $0.02, 0.03, 0.01, 0.03$ for hypotheses $H_0^{01}, H_0^{02}, H_0^{11}, H_0^{12}, H_0^{21}, H_0^{22}$, respectively. We therefore conclude that (7) is more appropriate to fit this data set than (8), and the improvement is statistically significant. Furthermore, as a referee pointed out, it is also of interest to test the linearity of the unknown coefficient functions in (7). For each $l=0,1,2$ and $s=1,2$, consider the null hypothesis that $\tilde{H}_0^{ls}:\alpha_{ls}(x_s)=\beta_{0,ls}+\beta_{1,ls}x_s$ for some coefficients $\beta_{0,ls}, \beta_{1,ls}$. The p-values are $0.016, <0.0001, 0.032, 0.025, 0.169, 0.021$ for hypotheses $\tilde{H}_0^{01}, \tilde{H}_0^{02}, \tilde{H}_0^{11}, \tilde{H}_0^{12}, \tilde{H}_0^{21}$,

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

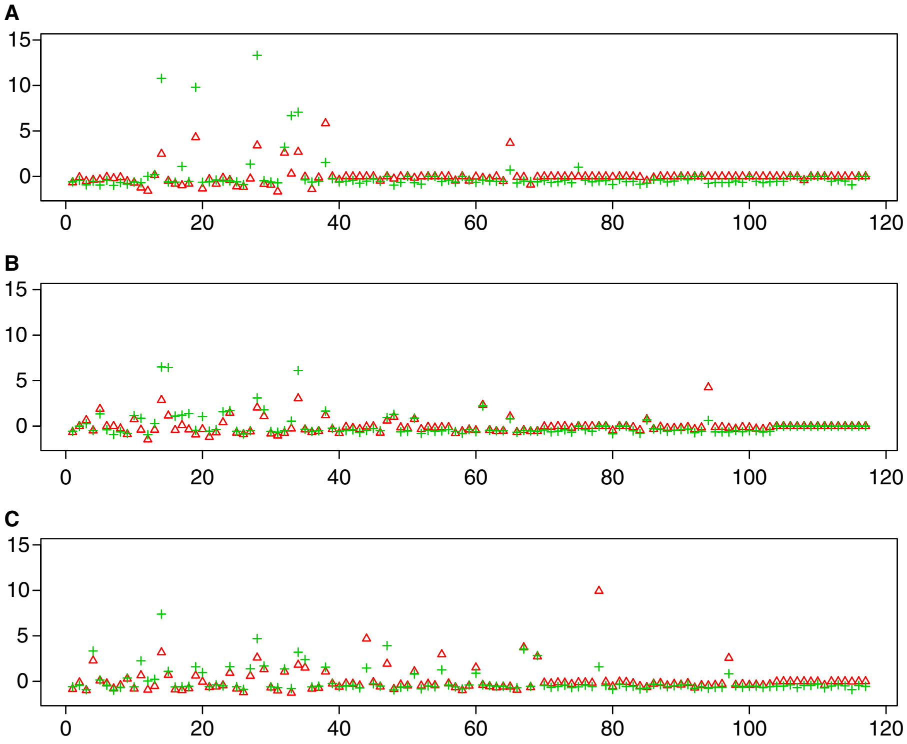

Fig. 6. The residual plots for the three periods: 1994/1995 (A), 1995/1996 (B), and 1996/1997 (C) are shown. In each plot, triangle and cross denote the residuals from (7) and (8), respectively.

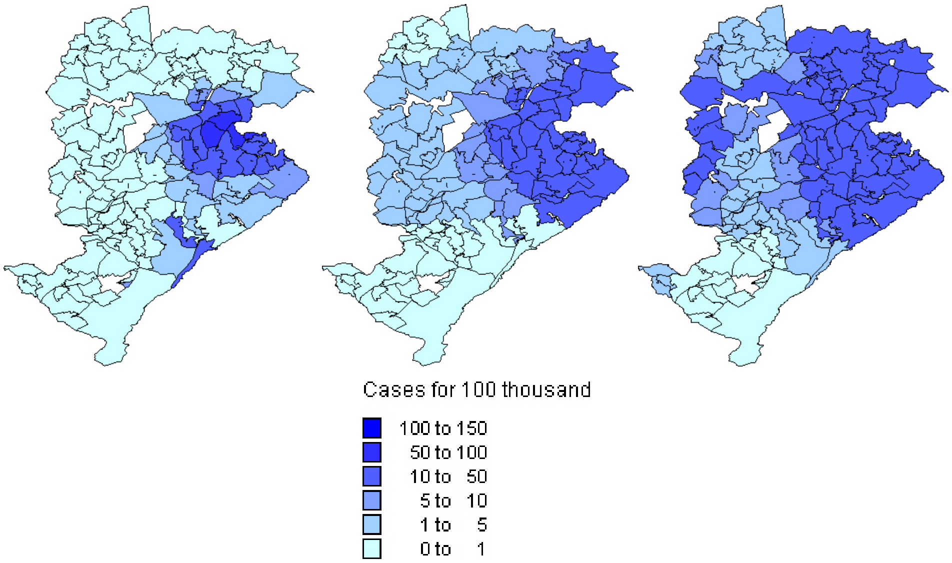

Fig. 7. The estimated human visceral Leishmaniasis rates (per 100 thousand) in the zones of Belo Horizonte using (7). The maps from left to right correspond to periods: 1994/1995, 1995/1996, and 1996/1997, respectively.

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

$\tilde{H}_0^{22}$, respectively. It suggests that only the coefficient function α_{21} in (7) is of linear form at significance level 0.05.

7. Conclusions

A polynomial spline estimation method together with a generalized likelihood ratio testing procedure have been proposed for the generalized semiparametric additive coefficient model. Theoretical results have been established under very broad assumptions on the data-generating process. Based on our experiences in working with both simulated and empirical examples, implementation of the proposed estimation method is as easy and fast as estimating a simple generalized linear model. The estimators' performance and their prediction power, however, are both promising as stipulated by theorem 1. These two aspects of the estimators, together with similar desirable properties of the generalized likelihood ratio testing procedure, make them highly recommendable for statistical inference in multivariate regression setting. A third feature, as mentioned in the introduction, is that the procedures of this paper automatically adapt to the generalized additive models (Hastie & Tibshirani, 1990; Härdle *et al.*, 2004), generalized varying coefficient models (Cai *et al.*, 2000), generalized partially linear models (Green & Silverman, 1994; Härdle *et al.*, 2000; Liang & Ren, 2005), generalized partially linear single index models (Carroll *et al.*, 1997), and simple generalized linear models. Hence, all these models can be simultaneously applied to any given data, and the most appropriate one can be selected via the generalized likelihood ratio testing procedure.

Acknowledgements

This research was partially supported by NIH/NIAID grants AI62247 and AI059773 and NSF grant DMS0806097. This paper improved greatly due to constructive comments from the editor and two referees. The authors thank Drs Renato M. Assunção and Ilka A. Reis for providing the data from the Brazil HVL study, and Drs David Birks, Jianhua Huang, and Lijian Yang for helpful discussions.

References

Assunção, R. (2003). Space varying coefficient models for small area data. *Environmetrics* **14**, 453–473.

Assunção, R., Reis, I. & Oliveira, C. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with Bayesian space-time model. *Statist. Med.* **20**, 2319–2335.

de Boor, C. (2001). *A practical guide to splines*. Springer, New York.

Cai, Z., Fan, J. & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888–902.

Carroll, R. J., Fan, J., Gijbels, I. & Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477–489.

Fan, J. & Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

Fan, J., Zhang, C. & Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153–193.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach*. Chapman & Hall, London.

Härdle, W., Liang, H. & Gao, J. (2000). *Partially linear models*. Physica-Verlag, Heidelberg.

Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer Verlag, Heidelberg.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall, London.

Hastie, T. & Tibshirani, R. J. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55**, 757–796.

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

Huang, J. Z. (1998). Functional ANOVA models for generalized regression. *J. Multivariate Anal.* **67**, 49–71.

Huang, J. Z. & Liu, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* **62**, 793–802.

Huang, J. Z., Kooperberg, C., Stone, C. J. & Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Ann. Statist.* **28**, 961–999.

Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111–128.

Liang, H. & Ren, H. (2005). Generalized partially linear measurement error models. *J. Comput. Graph. Statist.* **14**, 237–250.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall, London.

Pietrzykowski, T. (1972). A generalization of the potential method for conditional maxima on Banach, reflexive spaces. *Numer. Math.* **18**, 367–372.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590–606.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118–184.

Xue, L. & Yang, L. (2006a). Estimation of semiparametric additive coefficient model. *J. Statist. Plann. Inference* **136**, 2506–2534.

Xue, L. & Yang, L. (2006b). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16**, 1423–1446.

Received October 2008, in final form March 2009

Hua Liang, Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, New York 14642, USA.
E-mail: hliang@bst.rochester.edu

Appendix

A.1. Notation and assumptions

To formalize the discussion, we introduce two function spaces: the model space $\mathcal{M}$ and the spline approximation space $\mathcal{M}_n$. The model space $\mathcal{M}$ is a collection of functions on $\chi \times R^{d_1}$ defined as:

$$\mathcal{M}=\left\{m(\mathbf{x},\mathbf{t})=\sum_{l=1}^{d_1} a_l(\mathbf{x})t_l,\ \ a_l(\mathbf{x})=\alpha_{l0}+\sum_{s=1}^{d_2}\alpha_{ls}(x_s);\alpha_{ls}\in\mathcal{H}_s^0\right\},$$

where $\{\alpha_{l0}\}_{l=1}^{d_1}$ are finite constants, and $\mathcal{H}_s^0=\{\alpha(x):E\{\alpha(X_s)\}=0, E\{\alpha^2(X_s)\}<\infty\}$, for $1\leq s\leq d_2$. Then the predictor function $\eta(\mathbf{x},\mathbf{t})$ in (1) is modelled as an element of $\mathcal{M}$. Let $(Y_i,\mathbf{X}_i,\mathbf{T}_i)_{i=1}^n$ be a random sample of size n from the distribution of $(Y,\mathbf{X},\mathbf{T})$. In what follows, denote by E_n the empirical expectation. For functions $m_1,m_2\in\mathcal{M}$, define the theoretical inner product and the empirical inner product, respectively as:

$$\langle m_1,m_2\rangle = E\{m_1(\mathbf{X},\mathbf{T})m_2(\mathbf{X},\mathbf{T})\},$$

$$\langle m_1,m_2\rangle_n = E_n\{m_1(\mathbf{X},\mathbf{T})m_2(\mathbf{X},\mathbf{T})\}=\frac{1}{n}\sum_{i=1}^n m_1(\mathbf{X}_i,\mathbf{T}_i)m_2(\mathbf{X}_i,\mathbf{T}_i).$$

The induced norms are denoted as:

$$\|m_1\|^2=E\{m_1^2(\mathbf{X},\mathbf{T})\},\quad \|m_1\|_n^2=E_n\{m_1^2(\mathbf{X},\mathbf{T})\}. \tag{9}$$

We now define the polynomial spline approximation space $\mathcal{M}_n$,

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

$$\mathcal{M}_n = \left\{ m_n(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} g_l(\mathbf{x}) t_l, \quad g_l(\mathbf{x}) = c_{l0} + \sum_{s=1}^{d_2} g_{ls}(x_s); g_{ls} \in \varphi_s^{0,n} \right\},$$

in which $\varphi_s^{0,n} = \{g_s : g_s \in \varphi_s, \Sigma_{i=1}^n g_s(X_{is})/n = 0\}$, the empirically centred polynomial spline space. Lemma A.3 shows that the functions in $\mathcal{M}$ and $\mathcal{M}_n$ have essentially unique representations.

For any $m \in \mathcal{M}$, let $l(m(\mathbf{X}, \mathbf{T}), Y) = B[m(\mathbf{X}, \mathbf{T})]Y - C[m(\mathbf{X}, \mathbf{T})]$. When they exist, define the log-likelihood and the expected log-likelihood function separately as:

$$l_n(m) = \frac{1}{n} \sum_{i=1}^{n} l(m(\mathbf{X}_i, \mathbf{T}_i), Y_i), \quad \Lambda(m) = E[l(m(\mathbf{X}, \mathbf{T}), Y)].$$

Note that the expected log-likelihood function $\Lambda(\cdot)$ need not to be defined for all $m \in \mathcal{M}$. Therefore, we restrict our attention to a subset of $\mathcal{M}$, denoted as $\mathcal{M}^*$, which is a collection of bounded functions in $\mathcal{M}$. That is,

$$\mathcal{M}^* = \{m \in \mathcal{M} : \text{range}(m) \subset \mathcal{L}, \text{ for a compact subinterval } \mathcal{L} \subset \mathcal{R}\}.$$

Then $\Lambda(\cdot)$ is well defined on $\mathcal{M}^*$, under assumptions (C1)–(C5). The subinterval $\mathcal{L}$ is chosen to be large enough such that η and $2\eta \in \mathcal{M}^*$, where the true predictor function η is bounded on $\chi_1 \times \chi_2$, under assumptions (C3) and (C6). Similarly, one defines

$$\mathcal{M}_n^* = \{m_n \in \mathcal{M}_n : \text{range}(m_n) \subset \mathcal{L}, \text{ for a compact subinterval } \mathcal{L} \subset \mathcal{R}\}.$$

To prove the theoretical results, we need the following assumptions. In what follows, we have denoted by the same letters *c, C*, any positive constants without distinction in each case. For any funtion f on χ, denote $\|f\|_\infty = \sup_{x \in \chi} |f(x)|$, and denote by $C^p([0, 1])$, the space of p-times continuously differentiable functions on $[0, 1]$.

(C1) *The function $B(\cdot)$ is twice continuously differentiable and its first derivative $B'(\cdot)$ is strictly positive.*

(C2) *There is a subinterval S of $\mathcal{R}$, such that ρ is concentrated on S, and $B''(\eta)y - C''(\eta) < 0$, for all $\eta \in \mathcal{R}$ and $y \in S$.*

(C3) *The tuning variables $\mathbf{X} = (X_1, \ldots, X_{d_2})^{\mathrm{T}}$ and the linear covariates $\mathbf{T} = (T_1, \ldots, T_{d_1})^{\mathrm{T}}$ are compactly supported and without loss of generality, we assume that their supports are $\chi_1 = [0, 1]^{d_2}$, and $\chi_2 = [0, 1]^{d_1}$, respectively.*

(C4) *The joint density of $\mathbf{X}$, denoted by $f(\mathbf{x})$, is absolutely continuous and bounded away from zero and infinity, that is $0 < c \le \min_{\mathbf{x} \in \chi_1} f(\mathbf{x}) \le \max_{\mathbf{x} \in \chi_1} f(\mathbf{x}) \le C < \infty$.*

(C5) *Let $\lambda_0(\mathbf{x}) \le \cdots \le \lambda_{d_1}(\mathbf{x})$ be the eigenvalues of $E(\mathbf{T}\mathbf{T}^{\mathrm{T}} \mid \mathbf{X} = \mathbf{x})$. We assume that $\{\lambda_l(\mathbf{x})\}_{l=1}^{d_1}$ are uniformly bounded away from 0, and infinite, for all $\mathbf{x} \in \chi_1$.*

(C6) *The coefficient functions $\alpha_{ls} \in C^{p+1}([0, 1])$, for $l = 1, \ldots, d_1, s = 1, \ldots, d_2$.*

(C7) *There is a constant $C > 0$, such that $\sup_{\mathbf{x} \in \chi_1, \mathbf{t} \in \chi_2} \text{var}(Y \mid \mathbf{X} = \mathbf{x}, \mathbf{T} = \mathbf{t}) \le C$.*

(C8) *The d_2 sets of knots denoted as*

$$k_{s,n} = \{0 = x_{s,0} \le x_{s,1} \le \cdots \le x_{s,N_n} \le x_{s,N_n+1} = 1\}, \quad s = 1, \ldots, d_2,$$

are quasi-uniform, that is, there exists $c > 0$

$$\max_{s=1,\ldots,d_2} \frac{\max(x_{s,j+1} - x_{s,j}, j = 0, \ldots, N_n)}{\min(x_{s,j+1} - x_{s,j}, j = 0, \ldots, N_n)} \le c.$$

Furthermore, the number of interior knots $N_n \asymp n^{1/(2p+3)}$, where p denotes the degree of spline space and $\asymp$ denotes that both sides have the same order.

The assumption (C1) implies that the function $C(\cdot)$ is twice continuously differentiable, $A(\cdot)$ is continuously differentiable, and $A'(\cdot)$ is strictly positive. Furthermore, for each $\eta \in \mathcal{R}$, the function $B(\xi)A(\eta)-C(\xi)$ has a unique maximum at $\xi=\eta$. Thus, the function that maximizes $\Lambda(\cdot)$ is given by the true predictor function η. Let $h=\max_{s=1,\dots,d_2;j=0,\dots,Nn}|x_{s,j+1}-x_{s,j}|$. Then (C8) implies that $h\asymp n^{-1/(2p+3)}$.

The assumptions (C1)–(C8) are common in polynomial spline estimation literature. Assumptions (C1) and (C2) are the same as assumptions 1 and 2 of Huang (1998), and conditions on page 591 of Stone (1986). They are satisfied by many familiar exponential families including normal, binomial, Poisson, and Gamma distributions. The assumptions (C3)–(C5) and (C7) are similar to assumptions 1–4 in Huang *et al.* (2002). The assumptions (C6) and (C8) are also used in Xue & Yang (2006b).

A.2. Technical lemmas

The first three lemmas present properties of the spaces $\mathcal{M}$ and $\mathcal{M}_n$, which were proved in Xue & Yang (2006b) under a more general set-up.

Lemma A.1

Under assumptions (C3)–(C5)*, there exists a constant* $c>0$ *such that*

$$\|m\|_2^2\geq c\left\{\sum_{l=1}^{d_1}\left(\alpha_{l0}^2+\sum_{s=1}^{d_2}\|\alpha_{ls}\|_2^2\right)\right\}$$

for all $m=\sum_{l=1}^{d_1}\left(\alpha_{l0}+\sum_{s=1}^{d_2}\alpha_{ls}\right)t_l\in\mathcal{M}$.

Lemma A.2

Under assumptions (C3)–(C8)*, one has*

$$\sup_{g_1\in\mathcal{M}_n,g_2\in\mathcal{M}_n}\left|\frac{\langle g_1,g_2\rangle_n-\langle g_1,g_2\rangle}{\|g_1\|\|g_2\|}\right|=O_p\left(\sqrt{\frac{\log^2(n)}{nh}}\right).$$

In particular, there exists constants $0<c<1<C$*, such that except on an event whose probability tends to zero, as* $n\to\infty, c\|g\|\leq\|g\|_n\leq C\|g\|, \forall g\in\mathcal{M}_n$.

Lemma A.3

Under assumptions (C3)–(C8)*, the model space* $\mathcal{M}$ *is theoretically identifiable, i.e. for any* $m\in\mathcal{M}$, $\|m\|=0$*, implies* $m=0$ *a.s. The approximation space* $\mathcal{M}_n$ *is empirically identifiable, for any* $m_n\in\mathcal{M}_n$, $\|m_n\|_n=0$*, implies* $m_n=0$ *a.s.*

Lemma A.4

Under assumptions (C1)–(C6)*,* $\Lambda(\cdot)$ *is strictly concave over* $\mathcal{M}^*$*. That is, for any* $m_0,m_1\in\mathcal{M}^*$*, which are essentially different* (*different on a set of positive probability relative to the joint distribution of* $(\mathbf{X},\mathbf{T})$)*,*

$$\Lambda(m_0+t(m_1-m_0))>(1-t)\Lambda(m_0)+t\Lambda(m_1).$$

Proof. For any $m\in\mathcal{M}^*$, one has

$$\Lambda(m)=E\{B[m(\mathbf{X},\mathbf{T})]Y-C[m(\mathbf{X},\mathbf{T})]\}=E\{B[m(\mathbf{X},\mathbf{T})]A[\eta(\mathbf{X},\mathbf{T})]-C[m(\mathbf{X},\mathbf{T})]\}.$$

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

Note that $A[\eta(\mathbf{X},\mathbf{T})]\in S$ almost surely, as ρ is concentrated on S [assumption (C2)]. Therefore, assumption (C2) gives $B''[m(\mathbf{X},\mathbf{T})]A[\eta(\mathbf{X},\mathbf{T})]-C''[m(\mathbf{X},\mathbf{T})]<0$, almost surely, for any $m\in\mathcal{M}^*$. Let $\lambda[m,\mathbf{X},\mathbf{T}]=B[m(\mathbf{X},\mathbf{T})]A[\eta(\mathbf{X},\mathbf{T})]-C[m(\mathbf{X},\mathbf{T})]$ and $m^{(t)}=m_0+t(m_1-m_0)$. As a result, almost surely,

$$\lambda[m^{(t)},\mathbf{X},\mathbf{T}]>t\lambda[m_0,\mathbf{X},\mathbf{T}]+(1-t)\lambda[m_1,\mathbf{X},\mathbf{T}].$$

The lemma follows from this inequality.

Lemma A.5

Under assumptions (C1)–(C6)*, there exist positive numbers c_1 and c_2, such that for all $m\in\mathcal{M}^*$,*

$$-c_1\|m-\eta\|^2\le\Lambda(m)-\Lambda(\eta)\le -c_2\|m-\eta\|^2.$$

Proof. For any $m\in\mathcal{M}^*$. Let $m^{(t)}=(1-t)\eta+tm$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\Lambda(m^{(t)})\,|_{t=0}=0,$$

as η is the true predictor function. Hence, integration by parts gives

$$\Lambda(m)-\Lambda(\eta)=\int_0^1(1-t)\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Lambda(m^{(t)})\,\mathrm{d}t,$$

where

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Lambda(m^{(t)})=E\left\{(m(\mathbf{X},\mathbf{T})-\eta(\mathbf{X},\mathbf{T}))^2\left[B''(m^{(t)}(\mathbf{X},\mathbf{T}))A(\eta(\mathbf{X},\mathbf{T}))-C''(m^{(t)}(\mathbf{X},\mathbf{T}))\right]\right\}.$$

Assumptions (C1) and (C2) entail that there exist constants $c_1^*,c_2^*>0$ such that,

$$\inf\nolimits_{m\in\mathcal{M}^*}[B''(m)A(\eta)-C''(m^{(t)})]\ge -c_2^*,$$
$$\sup\nolimits_{m\in\mathcal{M}^*}[B''(m)A(\eta)-C''(m^{(t)})]\le -c_1^*.$$

The proof of the lemma is completed by taking $c_1=c_1^*/2$ and $c_2=c_2^*/2$. The proof is similar to those in lemma 4.1 of Huang (1998) and lemma 6 of Stone (1986).

A.3. Proof of theorem 1

Note that $\eta=\operatorname{argmax}_{m\in\mathcal{M}^*}\Lambda(m)$, and $\hat\eta=\operatorname{argmax}_{m\in\mathcal{M}_n}l_n(m)$. Write $\eta^*=\operatorname{argmax}_{m\in\mathcal{M}_n}\Lambda(m)$. Then, one has the following error decomposition:

$$\hat\eta-\eta=\hat\eta-\eta^*+\eta^*-\eta,$$

where $\hat\eta-\eta^*$ and $\eta^*-\eta$ can be understood as the estimation and approximation errors, respectively. Then the first part of theorem 1, $\|\hat\eta-\eta\|=O_p(h^{p+1}+1/\sqrt{nh})$ is obtained by showing that $\|\eta^*-\eta\|=O_p(h^{p+1})$, and $\|\hat\eta-\eta^*\|=O_p(1/\sqrt{nh})$, which is proved in steps 1 and 2, respectively. The rest of theorem 1 follows immediately from lemma A.1. The same idea of error decomposition is also used in Huang (1998) where the generalized functional ANOVA model is assumed instead.

Step 1 (the approximation error): Lemma A.5 entails that there exist constants $c_1,c_2>0$, such that for all $m\in\mathcal{M}^*$,

$$-c_1\|m-\eta\|^2\le\Lambda(m)-\Lambda(\eta)\le -c_2\|m-\eta\|^2. \quad (10)$$

Thus, for any constant $c > 0$, and any $m_n \in \mathcal{M}_n^* \subset \mathcal{M}^*$ with $\|m_n - \eta\| = ch^{p+1}$ (if such m_n exists), (10) entails that

$$\Lambda(m_n) - \Lambda(\eta) \le -c_2 c^2 h^{2(p+1)}. \tag{11}$$

On the other hand, the approximation theorem (de Boor, 2001) ensures that, there exist spline functions $g_{ls} \in \varphi_s^0$, and a constant $C > 0$ that do not depend on n, such that $\|g_{ls} - \alpha_{ls}\|_\infty \le Ch^{p+1}$, for $1 \le l \le d_1, 1 \le s \le d_2$. Let $m_n^* = \sum_{l=1}^{d_1} \left\{\alpha_{l0} + \sum_{s=1}^{d_2} g_{ls}(x_s)\right\} t_l \in \mathcal{M}_n^*$. By assumption (C3), one has

$$\|m_n^* - \eta\| \le \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|(g_{ls} - \alpha_{ls}) t_l\| \le \sum_{l=1}^{d_1} \sum_{s=1}^{d_2} \|g_{ls} - \alpha_{ls}\|_\infty \le c_3 h^{p+1},$$

in which $c_3 = d_1 d_2 C$. Thus (10) also gives

$$\Lambda(m_n^*) - \Lambda(\eta) \ge -c_1 c_3^2 h^{2(p+1)}. \tag{12}$$

By choosing c such that $c > c_3\sqrt{c_1/c_2}$, (11) and (12) entail that, when n is sufficiently large,

$$\Lambda(m_n) < \Lambda(m_n^*), \quad \text{for any } m_n \in S(c), \tag{13}$$

where $S(c) = \{m_n : m_n \in \mathcal{M}_n^*, \text{ with } \|m_n - \eta\| = ch^{p+1}\}$. Note that for such choice of c, $\|m_n^* - \eta\| \le c_3 h^{p+1} < ch^{p+1}$. Let $B(c) = \{m_n : m_n \in \mathcal{M}_n^*, \text{ with } \|m_n - \eta\| \le ch^{p+1}\}$, which is a closed bounded convex set in $\mathcal{M}_n^*$. Assumption (C1) and lemma A.5 entail that $\Lambda(\cdot)$ is a continuous concave functional on $\mathcal{M}_n^*$. Therefore, theorem 2 in Pietrzykowski (1972) ensures that $\Lambda(\cdot)$ has a maximum on $B(c)$. On the other hand, (13) ensures that the maximum must be in the interior of $B(c)$. Together with the concavity of $\Lambda(\cdot)$ and the definition of η^*, η^* exists, and satisfies $\|\eta^* - \eta\| < ch^{p+1}$, for n sufficiently large. Hence, $\|\eta^* - \eta\| = O_p(h^{p+1})$.

Step 2 (the estimation error): Let $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^{I_n}$ be an orthornormal basis for $\mathcal{M}_n$ with respect to the theoretical inner product, where $I_n = (1 + d_2 J_n) d_1$. Then one can write $\hat{\eta} = \Sigma_{j=1}^{I_n} \hat{\beta}_j \phi_j$, and $\eta^* = \sum_{j=1}^{I_n} \beta_j^* \phi_j$, for some coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{I_n})^{\mathrm{T}}$, and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{I_n}^*)^{\mathrm{T}}$. For any $m \in \mathcal{M}_n$, write $l_n(m) = l_n(\sum_{j=1}^{I_n} \beta_j \phi_j)$, as $l_n(\boldsymbol{\beta})$. Let $\mathbf{S}(\boldsymbol{\beta}) = (\partial/\partial\boldsymbol{\beta}) l_n(\boldsymbol{\beta})$ be the score at $\boldsymbol{\beta}$, which is an I_n-dimensional vector having entries

$$\frac{\partial}{\partial\beta_j} l_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \phi_j(\mathbf{X}_i, \mathbf{T}_i) \left\{B'[m(\mathbf{X}_i, \mathbf{T}_i)] Y_i - C'[m(\mathbf{X}_i, \mathbf{T}_i)]\right\},$$

and let $\mathbf{D}(\boldsymbol{\beta}) = (\partial^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathrm{T}}) l_n(\boldsymbol{\beta})$ be the $I_n \times I_n$ Hessian matrix, which has entries

$$\frac{\partial}{\partial\beta_j \partial\beta_{j'}} l_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \phi_j(\mathbf{X}_i, \mathbf{T}_i) \phi_{j'}(\mathbf{X}_i, \mathbf{T}_i) \left\{B''[m(\mathbf{X}_i, \mathbf{T}_i)] Y_i - C''[m(\mathbf{X}_i, \mathbf{T}_i)]\right\}.$$

Then, for any $m = \sum_{j=1}^{I_n} \beta_j \phi_j \in \mathcal{M}_n$ with $\{\phi_j\}_{j=1}^{I_n}$ being a set of orthornormal basis and $\|m - \eta^*\| = |\boldsymbol{\beta} - \boldsymbol{\beta}^*| = a/\sqrt{nh}$, for some constant $a > 0$ to be chosen later, Taylor expansion gives

$$\begin{aligned} l_n(\boldsymbol{\beta}) &= l_n(\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{S}(\boldsymbol{\beta}^*) \\ &\quad + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \left[\int_0^1 (1 - t) \mathbf{D}(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*)) \, \mathrm{d}t\right] (\boldsymbol{\beta} - \boldsymbol{\beta}^*). \end{aligned}$$

For any fixed $\varepsilon > 0$, lemma A.7 implies that one can choose a sufficiently large, such that $P(|\mathbf{S}(\boldsymbol{\beta}^*)| < ac_5/\sqrt{nh}) \ge 1 - \varepsilon$. Let $\mathcal{A} = \{|\mathbf{S}(\boldsymbol{\beta}^*)| < ac_5/\sqrt{nh}\}$. Then on event $\mathcal{A}$

$$|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}} \mathbf{S}(\boldsymbol{\beta}^*)| \le |\boldsymbol{\beta} - \boldsymbol{\beta}^*| |\mathbf{S}(\boldsymbol{\beta}^*)| < c_5 a^2/(nh). \tag{14}$$

Moreover, for such a, lemma A.8 implies that

$$(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^{\mathrm{T}}\left[\int_0^1 (1-t)\mathbf{D}(\boldsymbol{\beta}^*+t(\boldsymbol{\beta}-\boldsymbol{\beta}^*))\,\mathrm{d}t\right](\boldsymbol{\beta}-\boldsymbol{\beta}^*) \le -c_5|\boldsymbol{\beta}-\boldsymbol{\beta}^*|^2 = -c_5a^2/(nh) \tag{15}$$

except on an event whose probability tends to zero, as $n\to\infty$. Thus, (14) and (15) entail that, except on an event whose probability tends to zero, as $n\to\infty$, $l_n(\boldsymbol{\beta})<l_n(\boldsymbol{\beta}^*)$ for all $|\boldsymbol{\beta}-\boldsymbol{\beta}^*|=a/\sqrt{nh}$. Hence by the concavity of $l_n(\boldsymbol{\beta})$ (lemma A.6) and similar arguments as in step 1, $\hat\eta$ exists and satisfies $\|\hat\eta-\eta^*\|<a/\sqrt{nh}$. As ε is arbitrary, $\hat\eta$ exists except on an event whose probability tends to zero as $n\to\infty$, and satisfies $\|\hat\eta-\eta^*\|=O_p(1/\sqrt{nh})$.

In the following, we present the necessary lemmas used in the proof. The lemmas are presented here because they need notations introduced in the proof of theorem 1.

Lemma A.6

Under assumptions (C1)–(C8)*, there exists a* $c_4>0$*, such that, except on an event whose probability tends to zero as* $n\to\infty$*,*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}l_n[m_1+t(m_2-m_1)]\le -c_4\|m_2-m_1\|^2,$$

for $0<t<1$ *and all* $m_1,m_2\in\mathcal{M}_n^*$*. Thus the log-likelihood* $l_n(\cdot)$ *is strictly concave on* $\mathcal{M}_n^*$ *except on an event whose probability tends to zero as* $n\to\infty$*.*

Proof. Let $m_t=m_1+t(m_2-m_1)$, for $0<t<1$. It follows from assumptions (C1) and (C2) that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}l_n(m_t)=\frac{1}{n}\sum_{i=1}^n\Big\{[m_2(\mathbf{X}_i,\mathbf{T}_i)-m_1(\mathbf{X}_i,\mathbf{T}_i)]^2 \times\Big[B''[m_t(\mathbf{X}_i,\mathbf{T}_i)]Y_i-C''[m_t(\mathbf{X}_i,\mathbf{T}_i)]\Big]\Big\}.$$

Note that there is a constant $\delta>0$, such that $B''(\xi)y-C''(\xi)\le-\delta$, for all $\xi\in\mathcal{L}, y\in S$. Thus, the right-hand side of this equality is bounded by

$$-\frac{\delta}{n}\sum_{i=1}^n[m_2(\mathbf{X}_i,\mathbf{T}_i)-m_1(\mathbf{X}_i,\mathbf{T}_i)]^2=-\delta\|m_2-m_1\|_n^2\le -c\delta\|m_2-m_1\|^2,$$

by lemma A.2. The result follows by letting $c_4=c\delta$.

Lemma A.7

Under assumptions (C1)–(C8)*, for any constant* $c>0$*,*

$$\lim_{a\to\infty}\limsup_{n\to\infty}P\left(|\mathbf{S}(\boldsymbol{\beta}^*)|\ge ca/\sqrt{nh}\right)=0.$$

Proof. Note that $\boldsymbol{\beta}^*$ maximizes

$$\Lambda(\boldsymbol{\beta})=\Lambda\left(\sum_{j=1}^{I_n}\beta_j\phi_j\right)=E\left[B\left(\sum_{j=1}^{I_n}\beta_j\phi_j(\mathbf{X},\mathbf{T})\right)Y-C\left(\sum_{j=1}^{I_n}\beta_j\phi_j(\mathbf{X},\mathbf{T})\right)\right].$$

Thus, $\left.\dfrac{\partial}{\partial\boldsymbol{\beta}}\Lambda(\boldsymbol{\beta})\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}=0$, which implies that

$$E\left\{\phi_j(\mathbf{X},\mathbf{T})\left[B'(\eta^*)Y-C'(\eta^*)\right]\right\}=0,\quad \textit{for}\ \ 1\le j\le I_n.$$

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

Thus

$$E(|\mathbf{S}(\boldsymbol{\beta}^*)|^2) = \sum_{j=1}^{I_n} E\left[\frac{\partial}{\partial \beta_j} l_n(\boldsymbol{\beta}^*)\right]^2 = \frac{1}{n}\sum_{j=1}^{I_n} \operatorname{var}\left\{\phi_j(\mathbf{X},\mathbf{T})\left[B'(\eta^*)Y - C'(\eta^*)\right]\right\},$$

where

$$\begin{aligned}\operatorname{var}\{\phi_j(\mathbf{X},\mathbf{T})\left[B'(\eta^*)Y - C'(\eta^*)\right]\} &= E\left[\operatorname{var}\left\{\phi_j(\mathbf{X},\mathbf{T})\left[B'(\eta^*)Y - C'(\eta^*)\right] | \mathbf{X},\mathbf{T}\right\}\right]\\ &\quad + \operatorname{var}\left[E\left\{\phi_j(\mathbf{X},\mathbf{T})\left[B'(\eta^*)Y - C'(\eta^*)\right] | \mathbf{X},\mathbf{T}\right\}\right]\\ &\leq E\left[\phi_j^2(\mathbf{X},\mathbf{T})(B'(\eta^*))^2\sigma^2(\mathbf{X},\mathbf{T})\right]\\ &\leq cE\left[\phi_j^2(\mathbf{X},\mathbf{T})\right] = c\|\phi_j\|^2.\end{aligned}$$

Thus

$$E(|\mathbf{S}(\boldsymbol{\beta}^*)|^2) \leq \frac{c}{n}\sum_{j=1}^{I_n} E[\phi_j^2(\mathbf{X},\mathbf{T})] = \frac{c}{n}\sum_{j=1}^{I_n}\|\phi_j\|^2 = \frac{c[1+d_2(N_n+p)]d_1}{n} \asymp \frac{1}{nh},$$

and we complete the proof of the lemma.

Lemma A.8

Under assumptions (C1)–(C8)*, there exists a constant* $c_5 > 0$*, such that, for any fixed positive constant a, with probability approaching* 1*, as* $n \to \infty$*,*

$$(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\left[\int_0^1 (1-t)\mathbf{D}(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\,\mathrm{d}t\right](\boldsymbol{\beta} - \boldsymbol{\beta}^*) \leq -c_5|\boldsymbol{\beta} - \boldsymbol{\beta}^*|^2$$

for all $\boldsymbol{\beta} \in \mathcal{R}^n$*, with* $|\boldsymbol{\beta} - \boldsymbol{\eta}^*| = a/\sqrt{nh}$*.*

Proof. For any $m \in \mathcal{M}_n$ with $\|m - \eta^*\| = a/\sqrt{nh}$, lemma A.6 entails that there exists a constant $c_4 > 0$, such that, for $0 < t < 1$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} l_n(\eta^* + t(m - \eta^*)) \leq -c_4\|m - \eta^*\|^2,$$

except on an event whose probability tends to zero as $n \to \infty$. Also, note that

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} l_n(\eta^* + t(m - \eta^*)) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} l_n(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*))\\ &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\mathbf{D}(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*))(\boldsymbol{\beta} - \boldsymbol{\beta}^*).\end{aligned}$$

Thus, the lemma follows with $c_5 = c_4/\int_0^1 \left(1-t\right)\mathrm{d}t = c_4/2$.

A.4. Proof of theorem 2

We prove the result only when the null hypothesis $H_0 : \alpha_{ls} = 0$, for notation convenience. For higher-order polynomials, the proof follows similarly. Define the approximation space under H_0 as:

$$\begin{aligned}\mathcal{M}_n^0 = \Bigg\{ m_n(\mathbf{x},\mathbf{t}) = &\sum_{l' \neq l, l'=1}^{d_1}\left[\alpha_{l'0} + \sum_{s'=1}^{d_2} g_{l's'}(x_{s'})\right] t_{l'}\\ &+ \left[\alpha_{l0} + \sum_{s' \neq s, s'=1}^{d_2} g_{ls'}(x_{s'})\right] t_l; \quad g_{ls} \in \varphi_s^{0,n}\Bigg\},\end{aligned}$$

© 2009 Board of the Foundation of the Scandinavian Journal of Statistics.

which leaves out the spline approximation term for α_{ls}. Note that $\mathcal{M}_n^0 \subset \mathcal{M}_n$. Denote $\hat{\eta}^0 = \operatorname{argmax}_{m \in \mathcal{M}_n^0} l_n(m)$, the restricted maximum likelihood estimator of η under H_0. Recall that $\hat{\eta} = \operatorname{argmax}_{m \in \mathcal{M}_n} l_n(m)$, the unrestricted maximum likelihood estimator. Write $\hat{\eta}^0 = \sum_{j=1}^{I_n} \hat{\beta}_j^0 \phi_j$, and $\hat{\eta} = \sum_{j=1}^{I_n} \hat{\beta}_j \phi_j$, for some coefficients $\hat{\boldsymbol{\beta}}^0 = (\hat{\beta}_1^0, \ldots, \hat{\beta}_{I_n}^0)^{\mathrm{T}}$, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{I_n})^{\mathrm{T}}$. Then, the Taylor expansion gives

$$\begin{aligned} l_n(\hat{\eta}^0) - l_n(\hat{\eta}) = l_n(\hat{\boldsymbol{\beta}}^0) - l_n(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \mathbf{S}(\hat{\boldsymbol{\beta}}) \\ + (\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \left[\int_0^1 (1-t) \mathbf{D}(\hat{\boldsymbol{\beta}} + t(\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}))\, \mathrm{d}t \right] (\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}), \end{aligned}$$

in which $\mathbf{S}(\hat{\boldsymbol{\beta}}) = 0$ by the definition of maximum likelihood estimator, and

$$\begin{aligned} &(\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \left[\int_0^1 (1-t) \mathbf{D}(\hat{\boldsymbol{\beta}} + t(\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}))\, \mathrm{d}t \right] (\hat{\boldsymbol{\beta}}^0 - \hat{\boldsymbol{\beta}}) \\ &\quad = \frac{1}{n} \sum_{i=1}^n \left\{ \left[\hat{\eta}^0(\mathbf{X}_i, \mathbf{T}_i) - \hat{\eta}(\mathbf{X}_i, \mathbf{T}_i) \right]^2 \right. \\ &\quad \left. \times \int_0^1 (1-t) [B''(\eta_t(\mathbf{X}_i, \mathbf{T}_i)) Y_i - C''(\eta_t(\mathbf{X}_i, \mathbf{T}_i))]\, \mathrm{d}t \right\}, \end{aligned}$$

where $\eta_t = \hat{\eta} + t(\hat{\eta}^0 - \hat{\eta})$. By (C1) and (C2), there exist constants $0 < c < C$, such that $-C \leq B''(\xi) y - C''(\xi) \leq -c$, for all $\xi \in \mathcal{L}$, and $y \in R$. Also note that $\int_0^1 (1-t)\, \mathrm{d}t = 1/2$, and $\frac{1}{n} \sum_{i=1}^n \{\hat{\eta}^0(\mathbf{X}_i, \mathbf{T}_i) - \hat{\eta}(\mathbf{X}_i, \mathbf{T}_i)\}^2 = \|\hat{\eta}^0 - \hat{\eta}\|_n^2$. Therefore, $T_n = 2[l_n(\hat{\eta}) - l_n(\hat{\eta}^0)]$ satisfies $c\|\hat{\eta}^0 - \hat{\eta}\|_n^2 \leq T_n \leq C\|\hat{\eta}^0 - \hat{\eta}\|_n^2$. That is, $T_n \asymp \|\hat{\eta}^0 - \hat{\eta}\|_n^2$. Lemma A.2 further gives that $T_n \asymp \|\hat{\eta}^0 - \hat{\eta}\|^2$. Then, the result of theorem 2 is obtained by noting the following. Under H_0, $\|\hat{\eta}^0 - \hat{\eta}\| \leq \|\hat{\eta}^0 - \eta\| + \|\hat{\eta} - \eta\| = o_p(1)$. While under H_1, $\|\hat{\eta}^0 - \hat{\eta}\| \geq \|\hat{\eta}^0 - \eta\| - \|\hat{\eta} - \eta\|$, in which by lemma A.1, $\|\hat{\eta}^0 - \eta\|^2 \geq c\|\alpha_{ls}\|^2$, and $\|\hat{\eta} - \eta\| = o_p(1)$.