UCLA
Department of Statistics Papers

Title
Simultaneous Confidence Bands and Hypothesis testing in Varying-Coefficient Models

Permalink
https://escholarship.org/uc/item/0bh1c02p

Authors
Jianqing Fan
Wen-yang Zhang

Publication Date
2011-10-25

eScholarship.org Powered by the California Digital Library
University of California

Simultaneous Confidence Bands and Hypothesis Testing in Varying-Coefficient Models*

Jianqing Fan
Department of Statistics
University of California
Los Angeles, CA 90095-1554

Wenyang Zhang
Department of Statistics
The Chinese University of Hong Kong
Shatin, Hong Kong

Abstract

Regression analysis is one of the most commonly used techniques in statistics. When the dimension of independent variables is high, it is difficult to conduct efficient nonparametric analysis straightforwards from the data. As an important alternative to the additive and other nonparametric models, varying-coefficient models can reduce the modeling bias and avoid "curse of dimensionality" significantly. In addition, the coefficient functions can easily be estimated via a simple local regression. Based on local polynomial techniques, we provide the asymptotic distribution for the maximum of the normalized deviations of the estimated coefficient functions away from the true coefficient functions. Using this result and the pre-asymptotic substitution idea for estimating biases and variances, simultaneous confidence bands for the underlying coefficient functions are constructed. An important question in the varying coefficient models is if an estimated coefficient function is statistically significantly different from zero or a constant. Based on newly derived asymptotic theory, a formal procedure is proposed for testing whether a particular parametric form fits a given data set. Simulated and real-data examples are used to illustrate our techniques.

KEY WORDS: Varying-coefficient models, simultaneous confidence band, maximum deviation, bandwidth, bias, variance.

*Fan's research was partially supported by NSF Grant DMS-9803200. The authors are grateful to Professor S.Y. Lee for various helpful suggestions and comments.

1 Introduction

Regression analysis is one of the most commonly used techniques in statistics. The aim of the analysis is to explore the association between dependent and independent variables and to identify their impact on the dependent variable. If the mean response is linear, the linear regression technique is very useful. However, this assumption is restrictive, and it is not always granted. Motivated by various applications, many useful data-analytic modeling techniques have been proposed to extend the traditional parametric models; see for example the books by Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijbels (1996), among others. For high-dimensional regression analysis, it is difficult to make efficient statistical analysis straightforwards from the data without imposing some forms on the model. Many powerful approaches have been incorporated to avoid so-called "curse of dimensionality". Examples include additive modeling (Breiman and Friedman, 1995; Hastie and Tibshirani 1990), low-dimensional interaction modeling (Friedman 1991, Gu and Wahba, 1992, Stone *et al.* 1997), multiple-index models (Härdle and Stoker 1990, Li 1991), and partially linear models (Wahba 1984; Green and Silverman 1994), and their hybrids (Carroll *et al.* 1997), among others. An important alternative to the additive and other models is the varying-coefficient model (Cleveland, *et al.* 1991 and Hastie and Tibshirani, 1993), in which the coefficients of the linear models are replaced by smooth nonparametric functions and hence the regression coefficients are allowed to vary as functions of other factors. The varying-coefficient model is defined by the following linear model:

$$Y = \sum_{j=1}^{p} a_j(U) X_j + \varepsilon, \tag{1.1}$$

for given covariates $(U, X_1, \cdots, X_p)'$ and response variable Y with

$$E(\varepsilon | U, X_1, \cdots, X_p) = 0,$$

and

$$\mathrm{Var}(\varepsilon | U, X_1, \cdots, X_p) = \sigma^2(U).$$

By selecting $X_1 \equiv 1$, the model allows varying intercept in the model. Due to generality of the functions $a_j(U)$, the modeling bias can be reduced significantly while the "curse of dimensionality" is avoided. Moreover, it is well-recognized that (Hastie and Tibshirani, 1993 and its discussion), the model has wide applications. It is a useful extension of thresholding models in Tong (1990) and Chen and Tsay (1993) in the time series setup. It also appears natural in the longitudinal data analysis where one wishes to explore the extent to which covariates affect response changing over time. See for example Hoover *et al.* (1997), Wu, Chiang and Hoover (1998), and Fan and Zhang

(2000) for novel applications of the model to longitudinal data. The varying coefficient models are also useful for analyzing functional types of data. See Ramsay and Silverman (1997) and Brumback and Rice (1998) for details.

Assuming the coefficient functions $a_j(U)$ possess similar degree of smoothness, Hastie and Tibshirani (1993) proposed an estimate for $a_j(U)$ via the dynamic linear model (West *et al.* 1985; West and Harrison 1989) and an approach based on penalized least squares (Wahba 1990). Alternatively, local polynomial fitting is an attractive method both from theoretical and computation point of view. The idea of local polynomial regression has been around for a long time. It was systematically studied by Stone (1977, 1980) and Cleveland (1979). Recent work on local polynomial fitting includes Fan (1993), Ruppert and Wand (1994), Fan and Gijbels (1996), among others, in which a detailed picture is given on the merits of local polynomial fitting: the techniques have high minimax efficiency; they correct automatically excessive biases at boundary without using boundary kernels and adapt to various design points.

Based on local polynomial modeling, Fan and Zhang (1997) derived the asymptotic mean-square errors for the one-step and two-step procedures. In longitudinal data analysis, Wu *et al.* (1998) derived pointwise asymptotic normality for kernel smoothers and constructed confidence regions based on Bonferroni's adjustments. Zhang and Lee (1998) established the pointwise asymptotic normality of the estimated coefficient functions under model (1.1). Their results can be only used to construct the pointwise confidence intervals. This is unsatisfactory in many applications. For example, investigators often want to know if an estimated coefficient function is significantly away from zero or if an estimated coefficient function is really varying. This amounts to testing if the whole function is zero or constant. Hence a confidence band is needed for this case. In the present paper, we provide the asymptotic distributions for the maximum of the normalized deviations of the estimated coefficient functions from the true coefficient functions. The result is deeper than the pointwise asymptotic normality of Zhang and Lee (1998) and has other important statistical applications. With the maximum deviation result, one needs to estimate the bias and variance of the estimated coefficient functions. Our method is inspired by the pre-asymptotic substitution method of Fan and Gijbels (1995) in a nonparametric regression setup. It is demonstrated that the resulting simultaneous confidence intervals have approximate right coverage probability. This statement is also verified by our simulation studies.

An important inference question is if the coefficient function $a_j(U)$ in model (1.1) is zero or not so that one can assess if the variable X_j is statistically significant or not. The question can easily be answered by examining if the function zero is in the simultaneous confidence band or not. Further, one may naturally ask if a particular coefficient function $a_j(\cdot)$ is really varying. This

amounts to testing if $a_j(\cdot) = \theta$, an unknown constant. A semiparametric method of estimating θ is proposed and the asymptotic distribution of the test statistic is derived based on the maximum deviation. The idea is extended readily to testing a parametric null hypothesis: $a_j(u) = a_0(u, \theta)$, where $a_0(u, \theta)$ is a family of parametric models. While the testing statistic here is based on the maximum deviation, which appear natural in our context of constructing simultaneous confidence bands, other methods are also applicable in our context. In particular, the sieve likelihood method proposed recently by Fan *et al.* (1999) was demonstrated to be a powerful and wide-applicable nonparametric test. For various useful ideas of nonparametric hypothesis testing, see the recent book by Hart (1997).

The paper is organized as follows. In section 2, we briefly introduce the local polynomial fitting technique and its necessary notation. The asymptotic distributions of the normalized maximum deviations of the estimated coefficient functions from the true coefficient functions are derived. The result is used to construct simultaneous confidence intervals in Section 4, where techniques for estimating biases and variances are introduced. The simultaneous confidence bands are demonstrated to have right coverage probability asymptotically and empirically. In Section 5, we introduce various testing statistics for several hypothesis testing problems based on the maximum deviation theory. The null distributions of these testing statistics are derived. Section 6 analyzes an environmental data. Technical proofs are relegated to Section 7.

2 Estimation methods

Throughout this article, we assume that the coefficient functions $a_j(\cdot), \quad j = 1, \cdots, p$, in model (1.1) has $q+1$ continuous derivatives. Given a random sample $\{(U_i, X_{i1}, \cdots, X_{ip}, Y_i),\ i = 1, \cdots, n\}$ from model (1.1), the local polynomial modeling of order q will be adopted to estimate the coefficient functions $a_j(\cdot)$, $j = 1, \cdots, p$. It has been shown in Ruppert and Wand (1994) and Fan and Gijbels (1996) that local polynomial fits with odd orders outperform those with even orders. Hence, q is taken to be an odd integer. For each given point u_0, we approximate the function locally as

$$a_j(u) \approx \sum_{l=0}^{q} \frac{1}{l!} a_j^{(l)}(u_0)(u - u_0)^l, \tag{2.1}$$

for u in a neighborhood of u_0, where $a_j^{(l)}(\cdot)$ is the lth derivative of $a_j(\cdot)$. This leads to the following local least-squares problem: Minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{p} \sum_{k=0}^{q} c_{j,k}(U_i - u_0)^k X_{ij} \right\}^2 K_h(U_i - u_0), \tag{2.2}$$

with respect to $c_{j,k}$, $j = 1, \cdots, p$, $k = 0, \cdots, q$, for a given kernel function K and a bandwidth h, where $K_h(\cdot) = K(\cdot/h)/h$. Let

$$Y = (Y_1, \cdots, Y_n)^T, \qquad \mathbf{W} = \text{diag}\left(K_h(U_1 - u), \cdots, K_h(U_n - u)\right),$$

and

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{11}(U_1 - u)^q & \cdots & X_{1p} & \cdots & X_{1p}(U_1 - u)^q \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{n1}(U_n - u)^q & \cdots & X_{np} & \cdots & X_{np}(U_n - u)^q \end{pmatrix}.$$

From the solution to the least-squares problem (2.2), we obtain the estimator of $a_1(u)$ as:

$$\hat{a}_1(u) = e_{1,\kappa}^T \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} Y, \tag{2.3}$$

where $e_{i,j}$ denotes the unit vector of length j with 1 at position i, and $\kappa = p(q+1)$. Estimate for the other components can be obtained similarly.

The variance $\sigma^2(u)$ is a quantity that describes the noise level. Apart from the intrinsic interest as a parameter of the model, an estimator of this variance is essential in model selection and in carrying out statistical inferences about the coefficient functions, such as construction of confidence intervals and hypothesis testing. For $\sigma^2(u)$, we use the normalized weighted residual sum of squares from the local polynomial fit of order q to estimate it:

$$\hat{\sigma}^2(u) = \frac{1}{\text{tr}\left\{\mathbf{W} - \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W}^2 \mathbf{X}\right\}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 K_h(U_i - u), \tag{2.4}$$

where

$$\hat{Y} = (\hat{Y}_1, \cdots, \hat{Y}_n)^T = \mathbf{X} \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} Y.$$

Throughout this paper, we will use the following notation:

$$\mu_i = \int t^i K(t) dt, \quad \mathbf{u} = (\mu_{q+1}, \cdots, \mu_{2q+1})^T, \quad \text{and} \quad \nu_i = \int t^i K^2(t) dt.$$

Let Γ be a $(q+1) \times (q+1)$ matrix with the (i,j)th element μ_{i+j} and $\tilde{\Gamma}$ be the matrix similar to Γ except replacing μ_i by ν_i. Denote by $\mathcal{D}$ the observed covariates vector, namely

$$\mathcal{D} = (U_1, \cdots, U_n, X_{11}, \cdots, X_{1n}, \cdots, X_{p1}, \cdots, X_{pn})^T.$$

Set

$$\Omega(u) = E\left\{(X_1, \cdots, X_p)^T (X_1, \cdots, X_p) | U = u\right\}, \quad \boldsymbol{\epsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T,$$

With the above notation, we are ready to present our results.

3 Limiting distributions of maximum deviations

Without loss of generality, we assume that we are interested in constructing the simultaneous confidence intervals on the interval $[0, 1]$. For completeness, we present explicitly the following technical conditions.

(a) For an $s > 2$, $j = 1, \cdots, p$, $E|Y|^{2s} < \infty$, and $E|X_j|^{2s} < \infty$.

(b) The density function of U, $f(u)$, is continuous and positive on the interval $[0, 1]$.

(c) The matrix $\Omega(u)$ is non-singular and $\sigma^2(u) \neq 0$. Further, assume that the elements in $\Omega(u)$ and the function $\sigma(u)$ are continuous.

(d) For $j = 1, \cdots, p$, $E(X_j^{2s}|U = u)$ is bounded.

(e) The second derivative of $f(u)$ and the second derivative of $\sigma^2(u)$ are bounded.

(f) The kernel function $K(z)$ is a symmetric density function, and is absolutely continuous on its support set $[-A,\ A]$.

(f1) $K(A) \neq 0$ or

(f2) $K(A) = 0$, $K(z)$ is absolutely continuous and $K^2(z)$, $(K'(z))^2$ are integrable on the $(-\infty,\ +\infty)$.

(g) For $j = 1, \cdots, p$, $a_j^{(q+1)}(\cdot)$ is continuous.

For any function $g(u)$, define $\|g\|_\infty = \sup_{u \in [0,\ 1]} |g(u)|$ and for any matrix $A(u) = (a_{ij}(u))_p$, set

$$\|A\|_\infty = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} \|a_{ij}\|_\infty^2 \right)^{1/2}.$$

We first introduce a useful lemma which will be applied to prove our main result. It is interesting in its own right. Let $(U_1,\ \xi_1)$, $\cdots$, $(U_n,\ \xi_n)$ be independent and identically distributed random sample from $(U,\ \xi)$. We assume that U and the kernel function $K(\cdot)$ satisfy the above regularity conditions, and ξ satisfy

(a') for an $s > 2$, $E|\xi|^s < \infty$;

(b') the function $|r(u)|$ is bounded away from zero for $u \in [0,\ 1]$, and has a bounded first derivative on $[0, 1]$, where $r(u) = E(\xi^2|U = u)$;

(c') $\sup_x \int |y|^s f(x, y)dy = c_s < \infty$, where $f(x, y)$ is the joint density function of U and ξ.

Let

$$\mathbf{m}(u) = \frac{1}{\sqrt{nhf(u)r(u)}} \sum_{i=1}^{n} \xi_i K\left(\frac{U_i - u}{h}\right),$$

and

$$\mathbf{M}(u) = \mathbf{m}(u) - E\mathbf{m}(u).$$

For the process $\mathbf{M}(u)$, we have the following lemma:

Lemma 1 *Under assumptions (a')-(c') and (e), (f), if $h = n^{-b}$, for some $0 < b < 1 - 2/s$, we have*

$$P\left\{(-2\log h)^{1/2}\left(\nu_0^{-1/2}\|\mathbf{M}\|_\infty - d_n\right) < x\right\} \longrightarrow \exp\left\{-2e^{-x}\right\}$$

where with $\nu_0 = \int K^2(t)dt$,

$$d_n = (-2\log h)^{1/2} + \frac{1}{(-2\log h)^{1/2}}\left\{\log\frac{K^2(A)}{\nu_0\pi^{1/2}} + \frac{1}{2}\log\log h^{-1}\right\},$$

if assumption (f1) holds, and

$$d_n = (-2\log h)^{1/2} + \frac{1}{(-2\log h)^{1/2}}\log\left\{\frac{1}{4\nu_0\pi}\int (K'(t))^2 dt\right\},$$

if assumption (f2) is valid.

For kernel density estimation, Bickel and Rosenblatt (1973) obtained the asymptotic distribution of the maximum of the normalized deviation of the estimate from its expected value. Lemma 1 is a parallel result for regression. From Lemma 1, we can obtain

$$\sup_{u\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\left[K_h(U_i - u)\xi_i - E\left\{K_h(U_i - u)\xi_i\right\}\right]\right| = O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right),$$

which was obtained by Mack and Silverman (1982) under weaker conditions. See Lemma 2 in Section 7.

The proof of Lemma 1 can be obtained by the technique of the proof for Lemma 3 in Gruet (1996) and the technique in Bickel and Rosenblatt (1973).

From now on, let

$$\xi = e_{1,p}^T\Omega^{-1}(X_1, \cdots, X_p)^T\varepsilon, \quad r_1(u) = E(\xi^2|U = u) = e_{1,p}^T\Omega^{-1}(u)e_{1,p}\sigma^2(u),$$

$$\nu_{1,0} = \int K_1^2(t)dt, \quad K_1(t) = e_{1,q+1}^T\Gamma^{-1}(1, t, \cdots, t^q)^T K(t).$$

Note that the even positions of $e_{1,q+1}^T\Gamma^{-1}$ consist of zeros (see Fan, Gijbels, Hu, and Huang 1996). Thus, $K_1(t)$ is symmetric. Define the bias of an estimated coefficient as

$$\text{bias}(\hat{a}_j(u)|\mathcal{D}) = E(\hat{a}_j(u)|\mathcal{D}) - a_j(u).$$

We now give our main theorem as follows.

Theorem 1 *Under the assumptions (a)-(g) and $h = n^{-b}$, $1/(2q+3) \leq b < 1 - 2/s$, we have*

$$P\left\{(-2\log h)^{1/2}\left(\nu_{1,0}^{-1/2}\left\|\left(nhr_1^{-1}f\right)^{1/2}\left(\hat{a}_1 - a_1 - bias(\hat{a}_1|\mathcal{D})\right)\right\|_\infty - d_{v,n}\right) < x\right\}$$
$$\longrightarrow \exp\left\{-2e^{-x}\right\},$$

where $d_{v,n}$ is defined the same as d_n in Lemma 1, except ν_0 is now replaced by $\nu_{1,0}$ and $K(t)$ is replaced by $K_1(t)$.

Remark 1 *If the supremum in Theorem 1 is taken on an interval $[c,d]$ instead of [0,1], Theorem 1 continues to hold under suitable conditions, by using transformation arguments. The result reads as follows:*

$$P\left\{(-2\log\{h/(d-c)\})^{1/2}\left(\nu_{1,0}^{-1/2}\sup_{u\in[c,d]}\left|\left(nhr_1^{-1}f\right)^{1/2}\left(\hat{a}_1 - a_1 - bias(\hat{a}_1|\mathcal{D})\right)\right| - \tilde{d}_{v,n}\right) < x\right\}$$
$$\longrightarrow \exp\left\{-2e^{-x}\right\},$$

where $\tilde{d}_{v,n}$ is the same as $d_{v,n}$ in Theorem 1 except that h is now replaced by $h/(d-c)$.

4 Applications to constructing confidence bands

4.1 Confidence bands

Since $f(u)$, $\Omega(u)$, $\sigma^2(u)$ and $\mathrm{bias}(\hat{a}_1(u)|\mathcal{D})$ are unknown, Theorem 1 can not be directly used to construct the simultaneous confidence band for $a_1(u)$. Now, we give an estimating procedure for these unknown quantities.

By (2.3), the bias is given by

$$\mathrm{bias}(\hat{a}_1(u)|\mathcal{D}) = e_{1,\kappa}^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\beta},$$

where

$$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)^T, \quad \beta_i = \sum_{j=1}^{p}\left(a_j(U_i) - \sum_{k=0}^{q}\frac{1}{k!}a_j^{(k)}(u)(U_i-u)^k\right)X_{ij}.$$

Following the pre-asymptotic substitution method of Fan and Gijbels (1995), the bias can be approximated by

$$e_{1,\kappa}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\tau}$$

using the Taylor expansion, where the i-th element of the $n\times 1$ vector $\boldsymbol{\tau}$ equals to

$$\sum_{j=1}^{p}\left\{\frac{1}{(q+1)!}a_j^{(q+1)}(u)(U_i-u)^{q+1} + \frac{1}{(q+2)!}a_j^{(q+2)}(u)(U_i-u)^{q+2}\right\}X_{ij}.$$

Moreover, for $j = 1, \cdots, p$, by local polynomial fit of order $q+2$ with an appropriate pilot bandwidth h_* ($= O\left(n^{-1/(2q+5)}\right)$, which is optimal for estimating $a_j^{(q+1)}$), we can obtain the estimators $\hat{a}_j^{(q+1)}(u)$ and $\hat{a}_j^{(q+2)}(u)$. This gives an estimator of $\text{bias}(\hat{a}_1(u)|\mathcal{D})$ as follow:

$$\widehat{\text{bias}}(\hat{a}_1(u)|\mathcal{D}) = e_{1,\kappa}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\hat{\boldsymbol{\tau}},$$

where the ith element of the $n \times 1$ vector $\hat{\boldsymbol{\tau}}$ equals to

$$\sum_{j=1}^{p}\left\{\frac{1}{(q+1)!}\hat{a}_j^{(q+1)}(u)(U_i-u)^{q+1} + \frac{1}{(q+2)!}\hat{a}_j^{(q+2)}(u)(U_i-u)^{q+2}\right\}X_{ij}. \tag{4.1}$$

As for the unknown quantities $f(u)$, $\Omega(u)$ and $\sigma^2(u)$, a natural strategy is to estimate them separately, and then replace them by their corresponding estimators in Theorem 1. This strategy will involve too much asymptotic substitutions which are not elegant and accurate.

Except for a constant factor, $nhr_1^{-1}(u)f(u)$ in Theorem 1 is the asymptotic variance of $\hat{a}_1(u)$. Hence, it can be directly estimated from its pre-asymptotic counterpart. To see this, note that

$$\text{Var}(\hat{a}_1(u)|\mathcal{D}) = e_{1,\kappa}^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}\Psi\mathbf{W}\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}e_{1,\kappa},$$

where $\Psi = \text{diag}(\sigma^2(U_1), \cdots, \sigma^2(U_n))$. Let $\mathbf{G} = I_p \otimes \text{diag}(1, h, \cdots, h^q)$. Using Lemma 2 in section 7, and by a simple calculation, we have

$$\left\|\frac{h}{n}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}\Psi\mathbf{W}\mathbf{X}\mathbf{G}^{-1} - \sigma^2 f\Omega\otimes\tilde{\Gamma}\right\|_\infty = O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2} + h\right).$$

This together with (7.2) and (7.3) in section 7, using the properties of Kronecker product, by simple calculation, we obtain

$$\text{Var}(\hat{a}_1(u)|\mathcal{D}) = \frac{e_{1,q+1}^T\Gamma^{-1}\tilde{\Gamma}\Gamma^{-1}e_{1,q+1}}{nhf(u)}r_1(u)\left(1 + O_P\left(\left(\frac{\log(1/h)}{nh}\right)^{1/2} + h\right)\right), \tag{4.2}$$

uniformly for $u \in [0,\ 1]$. Note that

$$e_{1,q+1}^T\Gamma^{-1}\tilde{\Gamma}\Gamma^{-1}e_{1,q+1} = \nu_{1,0}.$$

Thus, the quantity $\nu_{1,0}\Big\{nhf(u)\Big\}^{-1}r_1(u)$ is naturally approximated by $\text{Var}(\hat{a}_1(u)|\mathcal{D})$.

The $\text{Var}(\hat{a}_1(u)|\mathcal{D})$ can be approximated by

$$e_{1,\kappa}^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}e_{1,\kappa}\sigma^2(u),$$

using the approximate local homoscedasticity. The unknown conditional variance $\sigma^2(u)$ can be estimated by using (2.4). To take advantage of the pilot estimation in assessing the bias, we can

estimate $\sigma^2(u)$ by the weighted residual sum of squares from a local $(q+2)$-order polynomial fit with bandwidth h_*, resulting in

$$\widehat{\mathrm{Var}}(\hat{a}_1(u)|\mathcal{D}) = e_{1,\kappa}^T \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} \left(\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right) \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} e_{1,\kappa}\hat{\sigma}^2(u). \tag{4.3}$$

This estimate is a by product from our pilot estimation of the bias. It involves less asymptotic substitution than separate estimation procedure and hence would expect to make better estimate of the variance. The above method can be justified by the following asymptotic result.

Theorem 2 *Under the conditions of Theorem 1, we have*

$$\begin{aligned} &P\left\{(-2\log h)^{1/2}\left(\left\|\left\{\mathit{Var}(\hat{a}_1|\mathcal{D})\right\}^{-1/2}\left(\hat{a}_1 - a_1 - \mathit{bias}(\hat{a}_1|\mathcal{D})\right)\right\|_\infty - d_{v,n}\right) < x\right\} \\ &\longrightarrow \exp\left\{-2e^{-x}\right\}. \end{aligned}$$

Furthermore, if $a_j^{(q+2)}(\cdot)$ and $\sigma''(\cdot)$ are continuous on $[0,1]$ and the pilot bandwidth h_* is of order $n^{-1/(2q+5)}$, then

$$\begin{aligned} &P\left\{(-2\log h)^{1/2}\left(\left\|\left\{\widehat{\mathit{Var}}(\hat{a}_1|\mathcal{D})\right\}^{-1/2}\left(\hat{a}_1 - a_1 - \widehat{\mathit{bias}}(\hat{a}_1|\mathcal{D})\right)\right\|_\infty - d_{v,n}\right) < x\right\} \\ &\longrightarrow \exp\left\{-2e^{-x}\right\}, \end{aligned}$$

provided that $nh^{2q+4}\log^3 h \to 0$.

Theorem 2 gives the following $1-\alpha$ confidence interval of $a_1(u)$ on $[0,\ 1]$:

$$\left(\hat{a}_1(u) - \widehat{\mathrm{bias}}(\hat{a}_1(u)|\mathcal{D}) - \Delta_{1,\alpha}(u),\ \hat{a}_1(u) - \widehat{\mathrm{bias}}(\hat{a}_1(u)|\mathcal{D}) + \Delta_{1,\alpha}(u)\right),$$

where

$$\Delta_{1,\alpha}(u) = \left(d_{v,n} + \left[\log 2 - \log\left\{-\log(1-\alpha)\right\}\right](-2\log h)^{-1/2}\right)\left\{\widehat{\mathrm{Var}}(\hat{a}_1(u)|\mathcal{D})\right\}^{1/2}.$$

That is, in the interval $[0,\ 1]$, the probability of the true curve $a_1(u)$ sandwiched between the curves $\hat{a}_1(u) - \widehat{\mathrm{bias}}(\hat{a}_1(u)|\mathcal{D}) - \Delta_{1,\alpha}(u)$ and $\hat{a}_1(u) - \widehat{\mathrm{bias}}(\hat{a}_1(u)|\mathcal{D}) + \Delta_{1,\alpha}(u)$ is $1-\alpha$. The confidence band for other components can be obtained similarly.

4.2 Simulations

In this section, we use two simulated examples to illustrate our method for constructing simultaneous confidence bands. The empirical coverage probabilities are observed. In both simulated models, we use model (1.1) with $p=2$, where X_1 and X_2 are normally distributed with correlation

coefficient $2^{-1/2}$, U follows a uniform distribution on $[0,\ 1]$ and $\varepsilon \sim N(0, \sigma^2)$. Further, U, ε and (X_1, X_2) are independent. The coefficient functions are taken as

$$\begin{aligned} &\text{Example 1:} \quad a_1(u) = \cos(2\pi u) \quad a_2(u) = 4u(1-u) \\ &\text{Example 2:} \quad a_2(u) = \sin(2\pi u) \quad a_2(u) = 4u(1-u). \end{aligned}$$

For each simulated example, we took $n = 500$ and noise to signal ratio about 1:5, namely

$$\sigma^2 = 0.2\text{Var}\{E(Y|U, X_1, X_2)\}.$$

The local linear fit with the Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ was used for estimating the regression coefficient functions. 95% simultaneous confidence bands are considered. The bandwidth is selected by a two-stage method based on residual squares criterion (RSC) (Zhang and Lee, 1998). The original idea of this method was proposed by Fan and Gijbels (1995). It is well known that it is difficult to estimate the bias since the bias involves high-order derivatives. Our modified method is to use the two-stage method to select the bandwidth $\hat{h}_{opt}$ for estimating the coefficient functions, and then use $0.5\hat{h}_{opt}$ as the bandwidth for constructing confidence bands. With this small bandwidth, we ignore the bias in the construction of simultaneous confidence band.

Table 1: *Coverage probabilities based on 500 simulations*

	$\alpha = 0.01$		$\alpha = 0.05$	
Coverage probabilities for	$a_1(u)$	$a_2(u)$	$a_1(u)$	$a_2(u)$
Example 1	0.99	0.99	0.91	0.94
Example 2	0.99	0.99	0.92	0.92

Table 1 summary the empirical coverage probabilities based on 500 simulations for $\alpha = 0.01$ and 0.05. The Monte Carol errors are of size $\sqrt{.95 \times .05/500} \approx 0.01$ for $\alpha = 0.05$. The coverage probabilities are quite close to the claimed confidence levels. Figure 1 depicts typical simultaneous confidence bands. The lengths of the simultaneous confidence bands are 3.17 times as large as the estimated standard errors.

5 Application to hypothesis testing

For varying-coefficient models, we often wish to know if an estimated coefficient function is significantly away from zero or if the estimated coefficient function is really varying. More generally, we

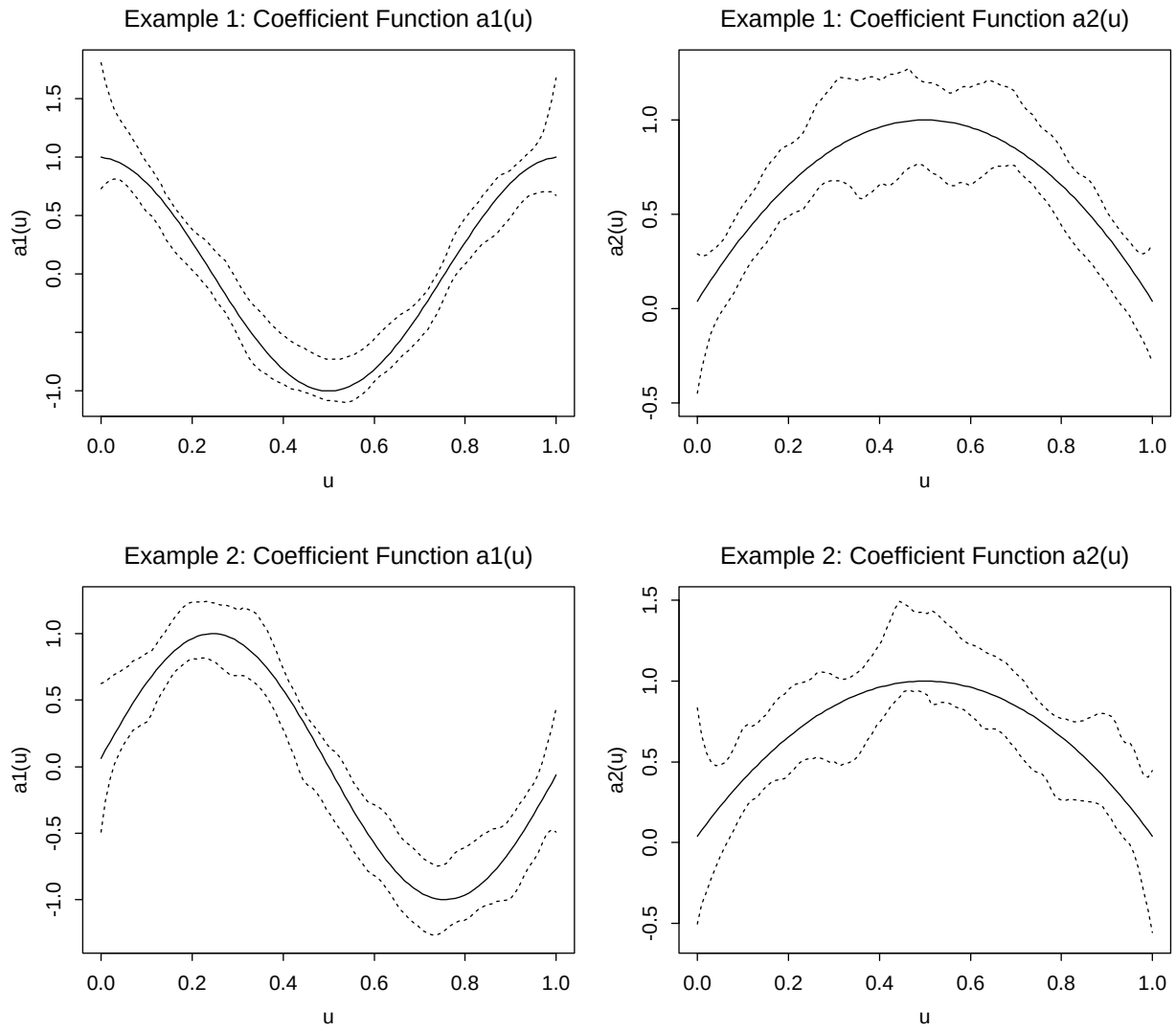

Figure 1: *Typical 95% simultaneous confidence bands for Example 1 (top panel) and Example 2 (bottom panel). The solid curves are the true curves, and the dashed curves are the* 95% *confidence bands.*

wish to test

$$H_0 : \ a_j(u) = a_0(u) \longleftrightarrow H_1 : a_j(u) \neq a_0(u)$$

for a given $a_0(\cdot)$. A natural test is to see if $a_0(\cdot)$ falls in the confidence band or not. This is equivalent to using the test statistic

$$(-2\log h)^{1/2}\left(\left\|\left\{\widehat{\mathrm{Var}}(\hat{a}_j|\mathcal{D})\right\}^{-1/2}\left(\hat{a}_j - a_0 - \widehat{\mathrm{bias}}(\hat{a}_j|\mathcal{D})\right)\right\|_\infty - d_{v,n}\right)$$

and rejecting H_0 when the test statistic exceeds the asymptotic critical value $c_\alpha = -\log(-0.5\log\alpha)$.

The above procedure extends readily to the composite null hypotheses. Consider the testing

problem:

$$H_0: \ a_j(u) = a_0(u,\theta) \longleftrightarrow H_1 : a_j(u) \neq a_0(u,\theta)$$

where θ is an unknown parameter. We first estimate θ via for example the least-squares method. If the estimator $\hat{\theta}$ of θ satisfies

$$\|a_0(\cdot,\hat{\theta}) - a_0(\cdot,\theta)\|_\infty = o_P\Big(h^{q+1}(\log h)^{-1}\Big),$$

which is usually true since $\hat{\theta}$ can be estimated at root-n rate, then we can apply Theorem 2 to find the null distribution of the test statistic

$$(-2\log h)^{1/2}\left(\left\|\left\{\widehat{\mathrm{Var}}(\hat{a}_j|\mathcal{D})\right\}^{-1/2}\Big(\hat{a}_j - a_0(\cdot,\hat{\theta}) - \widehat{\mathrm{bias}}(\hat{a}_j|\mathcal{D})\Big)\right\|_\infty - d_{v,n}\right).$$

This test amounts to checking if $a_0(\cdot,\hat{\theta})$ falls in the confidence band.

The following is devoted to testing if the coefficients are really varying. Without loss of generality, we consider the problem

$$H_0: \ a_p(u) = c \longleftrightarrow H_1 : a_p(u) \neq c. \tag{5.1}$$

Firstly, we propose a method to estimate c under the null hypotheses. Under the null hypotheses, the model is

$$Y = \sum_{j=1}^{p-1} a_j(U)X_j + cX_p + \varepsilon.$$

Set

$$\mathbf{X}_i = \begin{pmatrix} X_{11} & \cdots & X_{11}(U_1-U_i)^q & \cdots & X_{1p} & \cdots & X_{1p}(U_1-U_i)^q \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{n1}(U_n-U_i)^q & \cdots & X_{np} & \cdots & X_{np}(U_n-U_i)^q \end{pmatrix}$$

and

$$\mathbf{W}_i = \mathrm{diag}(K_h(U_1-U_i),\cdots,K_h(U_n-U_i)).$$

The estimate procedure for c consists of two steps. In the first step, we ignore the fact that c is a constant, and treat it as an unknown function $a_p(U)$. Based on local polynomial modeling, we obtain an estimator $\hat{a}_p(\cdot)$. Each of $\{\hat{a}_p(X_i)\}$ is an estimator of the unknown parameter c under the null hypothesis. In the second step, we average over these estimates to obtain a stabilized overall estimator:

$$\hat{c} = \frac{1}{n}\sum_{i=1}^n \hat{a}_p(U_i) = \frac{1}{n}\sum_{i=1}^n e_{\ell,\kappa}^T(\mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{W}_iY,$$

where $\ell = (p-1)(q+1)+1$. Here for simplicity of presentation, we assume that the design density U has support on $[0,1]$ and it is continuous and positive on $[0,1]$. Otherwise, a weighting scheme is needed so that only those $U_i's$ in the positive density regions are used. For the estimator $\hat{c}$, we have the following theorem.

Theorem 3 *Under the assumptions (a)-(d) and (g), if* $nh/\log h \longrightarrow \infty$, *then the conditional bias of* $\hat{c}$ *is*

$$bias(\hat{c}|\mathcal{D}) = O_P\Big(h^{q+2} + h^{q+1}\Big(-\log h/nh\Big)^{1/2}\Big),$$

and the conditional variance of $\hat{c}$ *is*

$$Var(\hat{c}|\mathcal{D}) = \frac{1}{n}\Big\{e_{1,q+1}^T\Gamma^{-1}(\mu_0,\cdots,\mu_q)^T\Big\}^2 E\Big\{e_{p,p}^T\Omega^{-1}(U)e_{p,p}\sigma^2(U)\Big\}\Big(1+o_P(1)\Big).$$

Based on Theorems 2 and 3, we test the problem (5.1) by computing the statistic

$$(-2\log h)^{1/2}\left(\left\|\Big\{\widehat{\mathrm{Var}}(\hat{a}_p|\mathcal{D})\Big\}^{-1/2}\Big(\hat{a}_p - \hat{c} - \widehat{\mathrm{bias}}(\hat{a}_p|\mathcal{D})\Big)\right\|_\infty - d_{v,n}\right)$$

and rejecting H_0 for large values of the test statistic.

6 Application to an environmental data set

We now illustrate the methodology via an application to an environmental data set. The data set consists of a collection of daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995 (Courtesy of Professor T.S. Lau). Of interest is to study the association between levels of pollutants and number of daily total hospital admissions for circulation and respiration and to examine the extent to which the association varies over time. We consider the relation among the number of daily hospital admission (Y) and level of pollutant Sulphur Dioxide X_2 (in $\mu g/m^3$), level of pollutant Nitrogen Dioxide X_3 (in $\mu g/m^3$), level of dust X_4 (in $\mu g/m^3$). We took $X_1 = 1$ — the intercept term, and $U = t =$ time. We first centered each of the three pollutants by their averages. For simplicity of notation, the resulting variables are still denoted as X_2, X_3 and X_4. The model

$$Y = a_1(t) + a_2(t)X_2 + a_3(t)X_3 + a_4(t)X_4 + \varepsilon$$

is used to fit the given data. Because of centering, the intercept can be interpreted as the expected number of admissions when pollutants are set at their averages. In our applications, the Epanechnikov kernel was employed and the bandwidths were chosen to be 20% of the interval length. The estimated coefficient functions along with the 95% simultaneous confidence bands were depicted in Figure 2.

A natural question is if these estimated coefficient functions are statistically significantly different from zero, and if so, whether they are really time varying. We now use our confidence-band method to answer this question. Table 2 presents the P-values for the test.

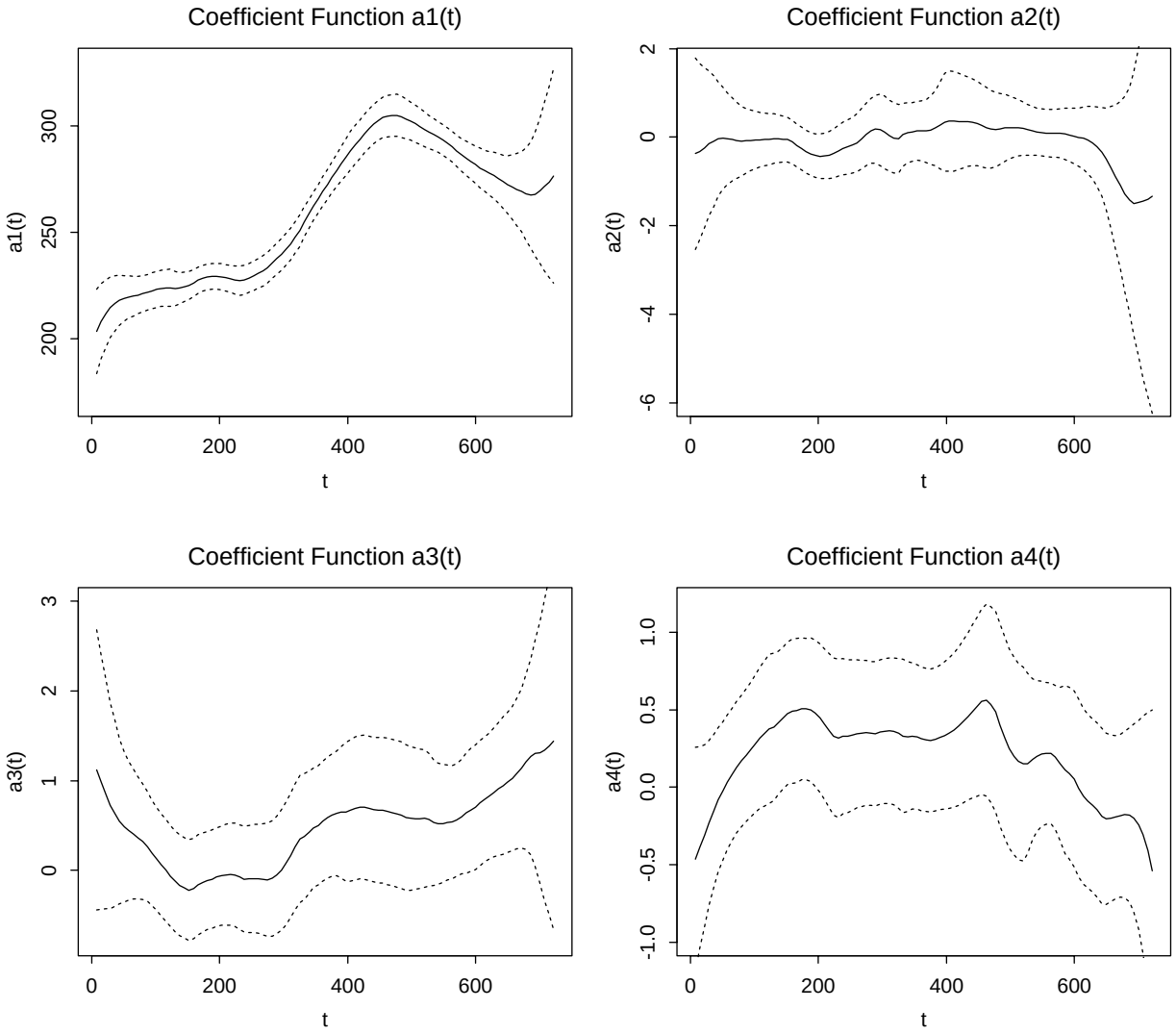

Figure 2: *95% simultaneous confidence bands for coefficient functions. The solid curves are the estimated coefficient functions, and the dashed curves are the* 95% *confidence bands.*

From Table 2, one can see that the covariate X_2, Sulphur Dioxide, is not significant at level 5%. In fact, from the Figure 2, the coefficient function for this variable is close to zero on the whole interval. All other variables are statistically significant, but the variable X_4, the level of dust, is not very highly significant. Indeed, by choosing the bandwidth to be 25% of the time interval under the study, the P-value for testing whether X_4 is significant or not becomes 0.048. The effects of intercept and Nitrogen Dixide X_3 are really varying with time, as shown in Table 2.

We now delete the insignificant variable X_2 and fit the varying coefficient model again. Figure 3 summarizes the result by plotting the intercept $a_1(t)$, which can be interpreted as the trend of hospital admissions (the expected number of admissions when the pollutants are set at their

Table 2: *P-values for testing if a coefficient function is zero or if a coefficient function is really time-varying*

Null hypothesis	$a_1(u)$	$a_2(u)$	$a_3(u)$	$a_4(u)$
$H_0 : a_j(\cdot) = 0$	0.0000	0.0999	0.0111	0.0287
$H_0 : a_j(\cdot) = c$	0.0000	0.3440	0.0135	0.0761

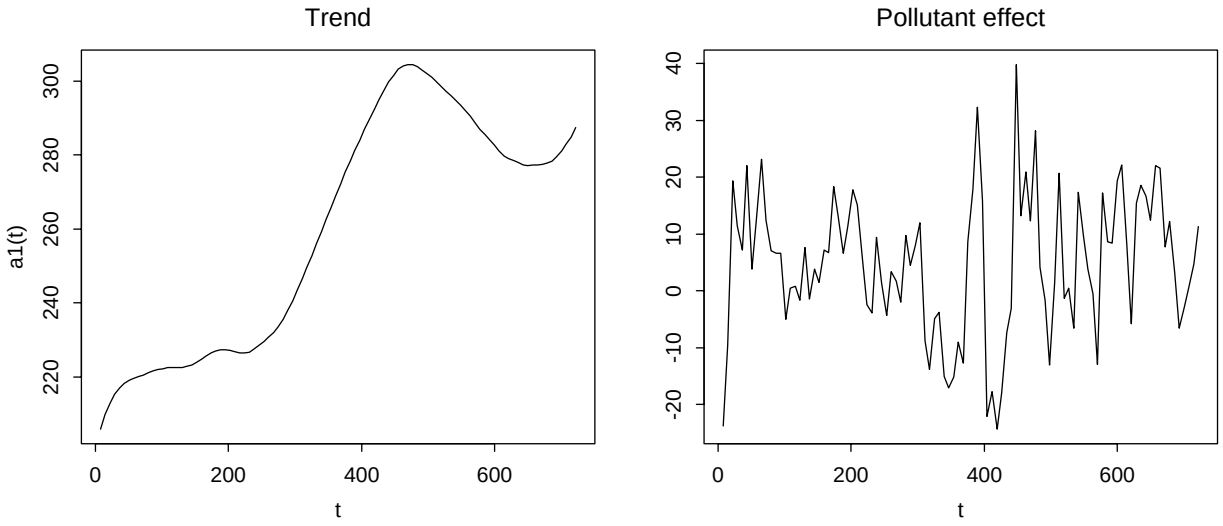

Figure 3: *The expected number of admissions are decomposed as the trend (left) and the contributions due to pollutants.*

averages) and the fitted function $\hat{a}_3(t)X_3(t) + \hat{a}_4(t)X_4(t)$, which can be understood as the expected number of hospital admissions due to pollution.

7 Proof

To obtain the proof of the theorems, the following lemma, which follows immediately from a result in Mack and Silverman(1982), is required.

Lemma 2 *Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be i.i.d random vectors, where Y_i's are scalar random variables. Assume further that $E|y|^s < \infty$ and $\sup_x \int |y|^s f(x, y)dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lips-*

chitz condition. Then

$$\sup_{x\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\left[K_h(X_i-x)Y_i-E\left\{K_h(X_i-x)Y_i\right\}\right]\right|=O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right),$$

provided that $n^{2\varepsilon-1}h\longrightarrow\infty$ *for some* $\varepsilon<1-s^{-1}$.

Obviously, Lemma 2 is also a corollary of Lemma 1, except different technical assumptions.

Proof of Theorem 1.

Obviously

$$\|\hat{a}_1-a_1-\mathrm{bias}(\hat{a}_1|\mathcal{D})\|_\infty=\sup_{u\in[0,1]}|e_{1,\kappa}^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon}|. \tag{7.1}$$

Let $\mathbf{I}_1(u)=e_{1,\kappa}^T\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon}$, and $\mathbf{G}=I_p\otimes\mathrm{diag}(1,h,\cdots,h^q)$. We approximate the process I_1 as follows.

First of all, we approximate the random matrix in I_1 by a deterministic one. By Lemma 2 and simple calculation, we have

$$\left\|\frac{1}{n}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{G}^{-1}-f\Omega\otimes\Gamma\right\|_\infty=O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}+h\right). \tag{7.2}$$

Using the fact

$$(A+hB)^{-1}=A^{-1}-hA^{-1}BA^{-1}+O\left(h^2\right)$$

and (7.2), we have

$$n\mathbf{G}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{G}=(f\Omega\otimes\Gamma)^{-1}+O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}+h\right), \tag{7.3}$$

uniformly for $u\in[0,1]$. By Lemma 7.2, we have

$$\left\|\frac{1}{n}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon}\right\|_\infty=O_P\left(\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}\right).$$

Using this and substituting (7.3) into I_1, we obtain

$$\left\|\mathbf{I}_1-\frac{1}{n}e_{1,\kappa}^T\left(f\Omega\otimes\Gamma\right)^{-1}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon}\right\|_\infty=O_P\left(h\left\{\frac{\log(1/h)}{nh}\right\}^{1/2}+\frac{\log(1/h)}{nh}\right). \tag{7.4}$$

Next, we consider the asymptotic distribution of

$$\mathbf{I}_2=\frac{1}{n}e_{1,\kappa}^T\left(f\Omega\otimes\Gamma\right)^{-1}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon}.$$

Let $K_{1,h}(t)=K_1(t/h)/h$. Then,

$$\mathbf{I}_2(u)=\frac{1}{nf(u)}\sum_{i=1}^{n}\xi_iK_{1,h}(U_i-u),$$

where $\xi_i = e_{1,p}^T \Omega^{-1}(X_{i1}, \cdots, X_{ip})^T \varepsilon_i$. By Lemma 1, we have

$$P\left\{(-2\log h)^{1/2}\left(\nu_{1,0}^{-1/2}\left\|\left(nhr_1^{-1}f\right)^{1/2}\mathbf{I}_2\right\|_\infty - d_{v,n}\right) < x\right\} \longrightarrow \exp\left\{-2e^{-x}\right\}.$$

This together with (7.1) and (7.4) yield the result of Theorem 1.

Proof of Theorem 2. First of all, using (4.2) and Theorem 1, we have

$$\begin{aligned}
&(-\log h)^{1/2}\left\|\left\{\mathrm{Var}(\hat{a}_1|\mathcal{D})\right\}^{-1/2}\left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D})\right)\right. \\
&\qquad\qquad \left. -\left\{nh(r_1\nu_{1,0})^{-1}f\right\}^{1/2}\left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D})\right)\right\|_\infty \\
= \;& \left\{(-\log h)^{1/2}\nu_{1,0}^{-1/2}\left\|(nhr_1^{-1}f)^{1/2}\left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D})\right)\right\|_\infty\right\}\left\|\left(\frac{r_1\nu_{1,0}}{nhf\,\mathrm{Var}(\hat{a}_1|\mathcal{D})}\right)^{1/2} - 1\right\|_\infty \\
= \;& o_P(1).
\end{aligned}$$

The first part of Theorem 2 follows from Theorem 1.

To prove the second part of Theorem 2, we first derive the rate of convergence for the bias and variance estimators. By using Lemma 2 and the standard arguments as in the proof of Theorem 1, we have

$$\|\widehat{\mathrm{bias}}(\hat{a}_1|\mathcal{D}) - \mathrm{bias}(\hat{a}_1|\mathcal{D})\|_\infty = O_P\Big(h^{q+1}(\log nh)^{1/2}\{n^{-2/(2q+5)}\log^{1/2} n + o(h)\}\Big), \tag{7.5}$$

where the rate $n^{-2/(2q+5)}\log^{1/2} n$ comes from the pilot estimation of $a_j^{(q+1)}(\cdot)$ and the term $o(h)$ comes from the coefficient in front of $\hat{a}_j^{(q+2)}(\cdot)$.

Next, we derive the rate of convergence for the variance estimator $\hat{\sigma}^2(\cdot)$. Let

$$c_n(u) = \mathrm{tr}\left\{\mathbf{W} - \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\right\}.$$

Recall $\hat{\sigma}^2(\cdot)$ was estimated from the pilot estimation with a local polynomial fit of order $(q+2)$ with the pilot bandwidth h_*. Then, using the notation $\boldsymbol{\beta}$ in Section 4.1 and the definition of $\hat{\sigma}^2(\cdot)$, we have

$$\begin{aligned}
\hat{\sigma}^2(u) &= c_n^{-1}(u)Y^T\Big(I - \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\Big)\mathbf{W}\Big(I - \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\Big)Y \\
&= c_n^{-1}(u)(\boldsymbol{\beta}+\boldsymbol{\epsilon})^T\left\{\mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\right\}(\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
&= c_n^{-1}(u)\boldsymbol{\epsilon}^T\left\{\mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\right\}\boldsymbol{\epsilon} \\
&\quad + c_n^{-1}(u)\boldsymbol{\beta}^T\left\{\mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\right\}\boldsymbol{\beta} \\
&\quad + 2c_n^{-1}(u)\boldsymbol{\epsilon}^T\left\{\mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\right\}\boldsymbol{\beta}.
\end{aligned}$$

By (7.2), (7.3) and Lemma 2, we have

$$\left\| c_n^{-1}\boldsymbol{\beta}^T \left\{ \mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W} \right\} \boldsymbol{\beta} \right\|_\infty = O_P\left(h_*^{2q+4}\right),$$

and

$$\left\| c_n^{-1}\boldsymbol{\epsilon}^T \left\{ \mathbf{W} - \mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W} \right\} \boldsymbol{\beta} \right\|_\infty = O_P\left(\left(\frac{-\log h_*}{nh_*} \right)^{1/2} h_*^{q+2} \right).$$

In addition, we note that

$$\left\| c_n^{-1}\boldsymbol{\epsilon}^T\mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\epsilon} \right\|_\infty = O_P\left(\frac{\log h_*}{nh_*} \right),$$

and

$$\| c_n^{-1}\boldsymbol{\epsilon}^T\mathbf{W}\boldsymbol{\epsilon} - \sigma^2 \|_\infty = O_P\left(\left(\frac{-\log h_*}{nh_*} \right)^{1/2} + h_*^2 \right).$$

Thus, using all of the above expressions, we get

$$\| \hat{\sigma}^2 - \sigma^2 \|_\infty = O_P\left(\left(\frac{-\log h_*}{nh_*} \right)^{1/2} + h_*^2 \right) = O_P\left(n^{-2/(2q+5)} \right). \tag{7.6}$$

Furthermore, using Lemma 2 and argued as in the proof of Theorem 1, we have

$$\left\| \frac{h}{n}\mathbf{G}^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\mathbf{G}^{-1} - f\Omega \otimes \tilde{\Gamma} \right\|_\infty = O_P\left(\left\{ \frac{\log(1/h)}{nh} \right\}^{1/2} + h \right).$$

This result together with (4.2), (4.3), (7.2), (7.3) and (7.6), by simple calculation, we have

$$\frac{\widehat{\mathrm{Var}}(\hat{a}_1(u)|\mathcal{D})}{\mathrm{Var}(\hat{a}_1(u)|\mathcal{D})} = 1 + O_P\left(\left(\frac{\log(1/h)}{nh} \right)^{1/2} + h + n^{-2/(2q+5)} \right),$$

uniformly for $u \in [0,1]$. Hence,

$$\begin{aligned}
& (-\log h)^{1/2} \left\| \left\{ \widehat{\mathrm{Var}}(\hat{a}_1|\mathcal{D}) \right\}^{-1/2} \left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D}) \right) \right. \\
& \qquad \left. - \left\{ \mathrm{Var}(\hat{a}_1|\mathcal{D}) \right\}^{-1/2} \left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D}) \right) \right\|_\infty \\
= \;& (-\log h)^{1/2} \left\| \left\{ \mathrm{Var}(\hat{a}_1|\mathcal{D}) \right\}^{-1/2} \left(\hat{a}_1 - a_1 - \mathrm{bias}(\hat{a}_1|\mathcal{D}) \right) \right\|_\infty \left\| \left\{ \mathrm{Var}(\hat{a}_1|\mathcal{D}) / \widehat{\mathrm{Var}}(\hat{a}_1|\mathcal{D}) \right\}^{1/2} - 1 \right\|_\infty \\
= \;& o_P(1).
\end{aligned}$$

This together with (7.5) leads to

$$P\left\{ (-2\log h)^{1/2} \left(\left\| \left\{ \widehat{\mathrm{Var}}(\hat{a}_1|\mathcal{D}) \right\}^{-1/2} \left(\hat{a}_1 - a_1 - \widehat{\mathrm{bias}}(\hat{a}_1|\mathcal{D}) \right) \right\|_\infty - d_{v,n} \right) < x \right\} \longrightarrow \exp\left\{ -2e^{-x} \right\}.$$

This completes the proof.

Proof of Theorem 3:

The conditional bias of $\hat{c}$ is

$$\begin{aligned}\mathrm{bias}(\hat{c}|\mathcal{D}) &= E\Big(\hat{a}_p(U_1)|\mathcal{D}\Big) - c \\ &= \frac{1}{(q+1)!} e_{\ell,\kappa}^T (\mathbf{X}_1^T \mathbf{W}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{W}_1 \times \\ &\quad \begin{pmatrix} 0 & \cdots & 0 \\ X_{21}(U_2-U_1)^{q+1} & \cdots & X_{2p}(U_2-U_1)^{q+1} \\ \vdots & \ddots & \vdots \\ X_{n1}(U_n-U_1)^{q+1} & \cdots & X_{np}(U_n-U_1)^{q+1} \end{pmatrix} \begin{pmatrix} a_1^{(q+1)}(U_1) \\ \vdots \\ a_{p-1}^{(q+1)}(U_1) \\ 0 \end{pmatrix} + O_P\Big(h^{q+2}\Big).\end{aligned}$$

Using Lemma 2, we have

$$\begin{aligned}&\frac{1}{n}\mathbf{G}^{-1}\mathbf{X}_1^T\mathbf{W}_1 \begin{pmatrix} 0 & \cdots & 0 \\ X_{21}(U_2-U_1)^{q+1} & \cdots & X_{2p}(U_2-U_1)^{q+1} \\ \vdots & \ddots & \vdots \\ X_{n1}(U_n-U_1)^{q+1} & \cdots & X_{np}(U_n-U_1)^{q+1} \end{pmatrix} \\ = &\quad \Big(f(U_1)\Omega(U_1) \otimes \mathbf{u} + O_P\Big(\Big(-\log h/nh\Big)^{1/2}\Big)\Big) h^{q+1}.\end{aligned}$$

This together with (7.2), and using the properties of Kronecker product, we obtain

$$\mathrm{bias}(\hat{c}|\mathcal{D}) = O_P\Big(h^{q+2} + h^{q+1}\Big(-\log h/nh\Big)^{1/2}\Big).$$

The conditional variance of $\hat{c}$ is

$$\begin{aligned}&\mathrm{Var}(\hat{c}|\mathcal{D}) \\ = &\quad \frac{1}{n^2}(1,\cdots,1) \begin{pmatrix} e_{\ell,\kappa}^T(\mathbf{X}_1^T\mathbf{W}_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{W}_1 \\ \vdots \\ e_{\ell,\kappa}^T(\mathbf{X}_n^T\mathbf{W}_n\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{W}_n \end{pmatrix} \Psi \\ &\qquad \times \begin{pmatrix} e_{\ell,\kappa}^T(\mathbf{X}_1^T\mathbf{W}_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{W}_1 \\ \vdots \\ e_{\ell,\kappa}^T(\mathbf{X}_n^T\mathbf{W}_n\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{W}_n \end{pmatrix}^T \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ = &\quad \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n e_{\ell,\kappa}^T(\mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_i^T\mathbf{W}_i\Psi\mathbf{W}_j\mathbf{X}_j(\mathbf{X}_j^T\mathbf{W}_j\mathbf{X}_j)^{-1}e_{\ell,\kappa}\end{aligned}$$

Using Lemma 2 and the properties of Kronecker product, and by some calculation, we have

$$\mathrm{Var}(\hat{c}|\mathcal{D}) = \frac{1}{n}\Big\{e_{1,q+1}^T\Gamma^{-1}(\mu_0,\cdots,\mu_q)^T\Big\}^2 E\Big\{e_{p,p}^T\Omega^{-1}(U)e_{p,p}\sigma^2(U)\Big\}\Big(1+o_P(1)\Big).$$

Hence, the result follows.

References

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. *Ann. Statist.*, **1**, 1071-1095.

Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.*, **93**, 961-994.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580–619.

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

Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.*, **88**, 298-308.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829-836.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991). Local regression models. In *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309–376. Wadsworth & Brooks, Pacific Grove.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, **21**, 196-216.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B.*, **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J., Gijbels, I., Hu. T.-C. and Huang, L.-S. (1996) An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation. *Statistica Sinica*, **6**, 113-127.

Fan, J. and Zhang, J.T. (2000). Functional linear models for longitudinal data. *Jour. Roy. Statist. Soc. B*, to appear.

Fan, J. and Zhang, W. (1997). Statistical estimation in varying coefficient models. *Technical report 230*, Department of Statistics, University of California at Los Angeles.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

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

Gruet, M.-A. (1996). A nonparametric calibration analysis *Ann. Statist.*, **24**, 1474-1492.

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". J. Comput. Graph. Statist. 2 (1993), 97–117.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, **84**, 986–995.

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

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

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer, New York.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, to appear.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, **86**, 316–342.

Mack, Y. P., Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **61**, 405–415.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*, Springer-Verlag, Berlin.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.*, **22**, 1346-1370.

Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595-645.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**, 1348-1360.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling. *Ann. Statist.*, **25**, 1371-1470.

Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329. Institute of Statistical Mathematics, Tokyo.

Wahba, G. (1990). *Spline Models for Observing Data.* Philadelphia: Society for Industrial and Applied Mathematics.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

West, M., Harrison, P. J. and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Amer. Statist. Assoc.*, **80**, 73-97.

West, M. and Harrison, P. J. (1989). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

Wu, C.O., Chiang, C.T. and Hoover, D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Amer. Statist. Assoc.*, **93**, 1388-1401.

Zhang, W. and S. Y. Lee. (1998). On local polynomial fitting of varying-coefficient models. Submitted for publication.