Journal of Nonparametric Statistics
Vol. 23, No. 2, June 2011, 367–383

Taylor & Francis
Taylor & Francis Group

Determination of linear components in additive models

Rong Chen[a], Hua Liang[b]* and Jing Wang[c]

[a]*Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA;* [b]*Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA;* [c]*Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA*

(*Received 30 June 2010; final version received 24 August 2010*)

Additive models have been widely used in nonparametric regression, mainly due to their ability to avoid the problem of the 'curse of dimensionality'. When some of the additive components are linear, the model can be further simplified and higher convergence rates can be achieved for the estimation of these linear components. In this paper, we propose a testing procedure for the determination of linear components in nonparametric additive models. We adopt the penalised spline approach for modelling the nonparametric functions, and the test is a sort of Chi-square test based on finite-order penalised spline estimators. The limiting behaviour of the test statistic is investigated. To obtain the critical values for finite sample problems, we use resampling techniques to establish a bootstrap test. The performance of the proposed tests is studied through simulation experiments and a real-data example.

Keywords: linearity test; partially linear additive models; penalised spline

1. Introduction

Nonparametric regression is a powerful tool for modelling the relationship between a response measurement Z and predictor measurements $X_1, \ldots, X_p$. Under the general setting that $Z = f(X_1, \ldots, X_p) + \varepsilon$, a nonparametric approach does not assume any specific functional form for the regression function $f(\cdot)$. Various nonparametric smoothing techniques have been developed and studied for the nonparametric estimation of the function f (Härdle, Müller, Sperlich, and Werwatz 2004; Li and Racine 2007). However, it has been found that the curse of dimensionality is one of the major problems that may arise in the estimation of high-dimensional functions (Stone 1986). Consequently, efforts have been made to develop methods that can reduce the dimensionality in nonparametric regression. See, for example, Hastie and Tibshirani (1990) for additive models; Hastie and Tibshirani (1993), Cai, Fan, and Li (2000), and Xia, Zhang, and Tong (2004) for varying-coefficient models, and Ichimura (1993) for single-index models. Among them, additive modelling is an effective strategy to reach this goal and has been extensively studied in the literature, including the backfitting algorithm (Hastie and Tibshirani 1990) and

*Corresponding author. Email: hliang@bst.rochester.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485252.2010.520713
http://www.informaworld.com

marginal integration (Linton and Nielsen 1995). Barry (1993), Chen, Liu, and Tsay (1995), Härdle and Tsybakov (1995), Sperlich, Tjøstheim, and Yang (2002), and Härdle, Huet, Mammen, and Sperlich (2004) have proposed methods to test for additivity, and thereby obtain diagnostic fools for using additive models in practice. These techniques have been widely used in applications and been implemented in general software packages such as Splus/R and Matlab. See Härdle, Müller et al. (2004) for a more detailed discussion on additive models.

Additive models assume that the conditional expectation of the response variable Z given $\mathbf{X} = (X_1, \ldots, X_p)$ is a sum of unknown one-dimensional functions of the individual variables and an intercept term, i.e.

$$E(Z|\mathbf{X}) = \alpha + \sum_{i=1}^{p} f_i(X_i).$$

One appealing feature of additive models is that nonparametric estimators of the unknown functions f_i have one-dimensional convergence rates (Stone 1986), which makes them much more accurate than estimating the p-dimensional function, and are able to avoid the 'curse of dimensionality'. This modelling strategy has been widely applied in biomedical, biological, epidemiological, and environmental research (Dominici, McDermott, Zeger, and Samet 2002; Cushman et al. 2004; Hsu et al. 2005).

Assuming that an additive structure is valid, it is often desirable to further simplify it to partial linear models (Härdle, Liang, and Gao 2000), which combine both linear components and nonparametric components in the additive setting. This is especially important when it is believed that the response variable Z depends on some independent variables linearly and others nonlinearly. Attractive features of partially linear models include that the linear component can be estimated at the root-n convergence rate, computation is easier for partially linear models than for additive models, and interpretation of the effect of each variable is easier and more intuitive than that of additive models.

In this paper, we consider the problem of determining linear components in additive models. Specifically, we try to answer the question that, under the pre-assumption of an additive model: $Z = \alpha + f_1(X_1) + \cdots + f_p(X_p) + \varepsilon$, how to determine (test) which function f_j is linear. Nonparametric linearity tests have been studied under univariate regression contexts by many researchers; e.g. Eubank and Speigelman (1990) and Härdle and Mammen (1993), among others. In the time series context, Hjellvik and Tjøstheim (1996) considered testing the linearity of a multivariate function using projections.

In this paper, we develop a testing procedure based on the square of the differences of fitted values under the null and alternative models. We estimate the nonparametric components using the penalised spline technique (Ruppert, Wand, and Carroll 2003) and study the issues of determining the subset of components in the additive model that are most likely to be linear. Asymptotic properties and smoothing parameter selection of a penalised spline estimator can be found in Wand (1999), Hall and Opsomer (2005), Li and Ruppert (2008), and Claeskens, Krivobokova, and Opsomer (2008). Penalised spline smoothing for the partially linear model in a more general setting has also been investigated lately, such as the single-index component estimation in additive models in Yu and Ruppert (2002), generalised additive models in Aerts, Claeskens, and Wand (2002), small area estimation in Opsomer, Claeskens, Ranalli, Kauermann, and Breidt (2008), and Spline-backfitted kernel smoothing for additive models in Wang and Yang (2007), Liu and Yang (2010), Ma and Yang (2011), and Song and Yang (2010).

The idea behind our test is similar to that of Hastie and Tibshirani (1990), where they proposed to use a likelihood-ratio test (LRT) and Chi-square approximations for the critical values. However, our approach is different in several ways. First, Hastie and Tibshirani (1990) focused on the backfitting algorithm by using smoothing spline or kernel regression. As a result, the selection penalty parameters or bandwidth is needed in each iteration, especially when the number of

smoothness of the functions f_i are substantially different. Hence, computational cost can be expensive in some cases and convergence is not always guaranteed. In our approach, we estimate the unknown functions using penalised splines, which avoids iteration. The resulting setting is a standard linear mixed-effects (LME) model. Computation implementation is straightforward. The penalty parameters are estimated through the restricted maximum-likelihood (RML) approach in an LME framework (see details below). The test statistic has an asymptotic weighted Chi-square distribution, similar to Crainiceanu, Ruppert, Claeskens, and Wand (2005), in which an exact LRT for penalised splines is used and the test statistics asymptotically follow a weighted Chi-square distribution. In addition, an approximate F-test is derived to compare the null and alternative models. This F-test has the same degrees of freedom as the F-test constructed to compare linear smoothers in Hastie and Tibshirani (1990).

The rest of the article is organised as follows. In Section 2, we start with an additive model with two regressors. We propose a statistic to test the linearity of one component and discuss its limiting distribution under the null hypothesis. A bootstrap version of the test is proposed as well and its limiting distribution is investigated. In Section 3, we generalise the testing procedure to a general additive model with multiple regressors. A step-wise testing procedure is also proposed for the determination of the linear components in the model. Section 4 presents simulation results that demonstrate the critical level and power of the proposed test. Section 5 analyses a real data set using our method. Final remarks are included in Section 6. All technical proofs are shown in the appendix.

2. The test statistic for a two-regressor additive model

In this section, we consider an additive model with two regressors and test if one of the components is linear. In the next section, we will study more general settings. Let

$$Z = f(X) + g(Y) + \varepsilon, \tag{1}$$

where X and Y are both explanatory variables, and $f(\cdot)$ and $g(\cdot)$ are unknown functions in $\Re$, with $g(0) = 0$. Here ε has a zero mean and finite variance and is independent of (X, Y). We also assume that both X and Y have finite support, without loss of generality, on $[0, 1]$.

2.1. *Penalised estimation*

First, we estimate the nonparametric functions f and g using the penalised spline techniques. To test if f is a linear function, we derive a test statistic based on the difference of the fitted values under the null hypothesis (f is linear) and the alternative (f is not linear). The penalised spline method for nonparametric regression was first proposed by Wahba (1990), and deeply studied by Eilers and Marx (1996) and Ruppert and Carroll (2000). See Ruppert et al. (2003) and the references therein for a comprehensive review. The approach under our setting can be briefly described as follows.

Assume the data (X_i, Y_i, Z_i) are iid samples of size n generated from Equation (1). We approximate f and g by $f(x; \tilde{\boldsymbol{\beta}}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \sum_{k=1}^{K_x} \beta_{2+k}(x - \zeta_k)_+^2$ and $g(y; \tilde{\boldsymbol{\gamma}}) = \gamma_1 y + \gamma_2 y^2 + \sum_{k=1}^{K_y} \gamma_{2+k}(y - \eta_k)_+^2$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2; \beta_3, \ldots, \beta_{2+K_x})^{\mathrm{T}}$, $\tilde{\boldsymbol{\gamma}} = (\gamma_1, \gamma_2; \gamma_3, \ldots, \gamma_{2+K_y})^{\mathrm{T}}$, $\zeta_1 < \cdots < \zeta_{K_x}$, and $\eta_1 < \cdots < \eta_{K_y}$ are fixed knots, and $a_+ = \max(a, 0)$.

The estimators of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ are defined as the minimisers of

$$\sum_{i=1}^{n} \{Z_i - f(X_i; \boldsymbol{\beta}) - g(Y_i; \boldsymbol{\gamma})\}^2 + \alpha_x \sum_{k=1}^{K_x} \beta_{2+k}^2 + \alpha_y \sum_{k=1}^{K_y} \gamma_{2+k}^2, \tag{2}$$

where α_x and α_y are two penalty parameters. In what follows, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ stand for $(\beta_0, \beta_1, \beta_2)$ and (γ_1, γ_2), respectively. The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\beta}$ and γ based on Equation (2) are equivalent to the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ based on the following LME model:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Y}\boldsymbol{\gamma} + \boldsymbol{\Xi}(x)\boldsymbol{b} + \boldsymbol{\Xi}(y)\boldsymbol{d} + \boldsymbol{\varepsilon}, \tag{3}$$

where

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 & Y_1^2 \\ Y_2 & Y_2^2 \\ \vdots & \vdots \\ Y_n & Y_n^2 \end{pmatrix}, \quad \boldsymbol{\Xi}(x) = \begin{Bmatrix} (X_1 - \zeta_1)_+^2 & \cdots & (X_1 - \zeta_{K_x})_+^2 \\ (X_2 - \zeta_1)_+^2 & \cdots & (X_2 - \zeta_{K_x})_+^2 \\ \vdots & \ddots & \vdots \\ (X_n - \zeta_1)_+^2 & \cdots & (X_n - \zeta_{K_x})_+^2 \end{Bmatrix},$$

$$\boldsymbol{\Xi}(y) = \begin{Bmatrix} (Y_1 - \eta_1)_+^2 & \cdots & (Y_1 - \eta_{K_y})_+^2 \\ (Y_2 - \eta_1)_+^2 & \cdots & (Y_2 - \eta_{K_y})_+^2 \\ \vdots & \ddots & \vdots \\ (Y_n - \eta_1)_+^2 & \cdots & (Y_n - \eta_{K_y})_+^2 \end{Bmatrix}, \quad \boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^{\mathrm{T}}.$$

And $\boldsymbol{b} = (b_1, \ldots, b_{K_x})^{\mathrm{T}} \sim N(0, \sigma_b^2)$, $\boldsymbol{d} = (d_1, \ldots, d_{K_y})^{\mathrm{T}} \sim N(0, \sigma_d^2)$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \sim N(0, \sigma_\varepsilon^2)$, and $\alpha_x = \sigma_\varepsilon^2/\sigma_b^2$, $\alpha_y = \sigma_\varepsilon^2/\sigma_d^2$.

This fact implies that the penalised spline smoother under the framework in Equation (2) is equivalent to a standard LME (Brumback, Ruppert, and Wand 1999; Crainiceanu et al. 2005). The solution can be obtained readily, such as through the use of an LME macro available for S-PLUS software. The penalty parameters α_x and α_y can be estimated as $\hat{\alpha}_x = \hat{\sigma}_\varepsilon^2/\hat{\sigma}_b^2$ and $\hat{\alpha}_y = \hat{\sigma}_\varepsilon^2/\hat{\sigma}_d^2$ through an RML approach. Note that in the general penalised spline framework, the penalty parameters and the number of knots K must be selected. For these two parameters, the penalty parameters play a more important role when compared with the number of knots. See Ruppert (2002) for a detailed discussion of selection of these parameters. Because the formulation of a mixed-effects model automatically estimates the penalty parameters, only the number of knots is needed to be specified beforehand. Our experience indicates $\max(10, n/4)$ is a good choice for the number of knots and that the results are insensitive to its choice. We also choose the knot locations at equally spaced sample quantiles of $\{X_i\}$ and $\{Y_i\}$, respectively.

In an analogous way, the partial linear model in which f is linear can be written as

$$\boldsymbol{Z} = \boldsymbol{X}_{[1,2]}\boldsymbol{\beta}_{[1,2]} + \boldsymbol{Y}\boldsymbol{\gamma} + \boldsymbol{\Xi}(y)\boldsymbol{d} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{X}_{[1,2]}$ is an $n \times 2$ matrix comprising by the first two columns of $\boldsymbol{X}$. We may consider Equation (4) as the model under the null hypothesis and Equation (3) as that under the alternative.

2.2. *Asymptotic tests*

Let $Z_{a,i}$ and $Z_{0,i}$ be the fitted values based on Equations (3) and (4). We use the L_2 distance between the two sets of fitted values as the test statistic:

$$\mathcal{T}_n = \sum_{i=1}^{n} (Z_{a,i} - Z_{0,i})^2. \tag{5}$$

For $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ based on either Equation (3) or (4), the corresponding fitted values can be always written in the vector-matrix notation as $\hat{\mathbf{Z}}_a = A_a \cdot \mathbf{Z}$ and $\hat{\mathbf{Z}}_0 = A_0 \cdot \mathbf{Z}$, where A_a and A_0 are $n \times n$ symmetric matrices and $A_a = C_a(C_a^{\mathrm{T}} C_a + B_a)^{-1} C_a^{\mathrm{T}}$, $A_0 = C_0(C_0^{\mathrm{T}} C_0 + B_0)^{-1} C_0^{\mathrm{T}}$, with $C_a = [\boldsymbol{X}\, \boldsymbol{Y}\, \boldsymbol{\Xi}(x)\ \boldsymbol{\Xi}(y)]$, $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 I_n$,

$$G_a = \mathrm{cov}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{d} \end{pmatrix} = \begin{bmatrix} \sigma_b^2 I_{K_x} & 0 \\ 0 & \sigma_d^2 I_{K_y} \end{bmatrix}, \quad B_a = \sigma_\varepsilon^2 \begin{bmatrix} 0 & 0 \\ 0 & G_a^{-1} \end{bmatrix},$$

and C_0, B_0, G_0 defined similarly. The test statistics $\mathcal{T}_n$ can then be represented as

$$\mathcal{T}_n = \mathbf{Z}^{\mathrm{T}}(A_a - A_0)^2 \mathbf{Z}.$$

It is noted that if $\boldsymbol{Z}$ is replaced by $f(\boldsymbol{x}) + g(\boldsymbol{y}) + \boldsymbol{\varepsilon}$, several terms in the expansion of the above expression would disappear asymptotically, and the expression can be asymptotically simplified to $\boldsymbol{\varepsilon}^{\mathrm{T}}(A_a - A_0)^2 \boldsymbol{\varepsilon}$. The next result shows an approximation to the distribution of the test statistic $\mathcal{T}_n$.

THEOREM 2.1 *Under Assumptions* (A1)–(A3) *listed in the appendix, we have*

$$\mathcal{T}_n = \mathbf{Z}^{\mathrm{T}}(A_a - A_0)^2 \mathbf{Z} \sim \mathbf{c}\mathcal{X}^2(b)$$

approximately, in the sense that the first two moments of $\mathcal{T}_n$ *are the same as the* $\mathcal{X}^2(b)$ *on the right-hand side, as n goes to infinity. Here*

$$\mathbf{c} = \sigma_\varepsilon^2 \frac{\sum_{i=1}^{K'} \lambda_i^2}{\sum_{i=1}^{K'} \lambda_i}, \quad b = \frac{\left(\sum_{i=1}^{K'} \lambda_i\right)^2}{\sum_{i=1}^{K'} \lambda_i^2}, \tag{6}$$

$\lambda_i' s$ *are the eigenvalues of the matrix* $(A_a - A_0)^2$*, and* K' *is the rank of* $(A_a - A_0)^2$.

The proof is given in the appendix. Given $\mathbf{c}$ and b, the critical value of an α-level test can be obtained based on the weighted Chi-square distribution.

An alternative method is to use the F-test, similar to that of Hastie and Tibshirani (1990). Let RSS_a and RSS_0 be the residual sums of squares under the alternative and null hypotheses and define $\Lambda_0 = (A_0 - I_n)^2$, $\Lambda_a = (A_a - I_n)^2$. If $\hat{Z}_a$ is unbiased, and $\hat{Z}_0$ is unbiased under the null hypothesis, we have the following theorem.

THEOREM 2.2 *Under Assumptions* (A1)–(A3), *asymptotically*

$$\frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/(\gamma_a - \gamma_0)}{\mathrm{RSS}_a/(n - \gamma_a)} \sim F_{\nu_1^2/\nu_2, \delta_1^2/\delta_2}, \tag{7}$$

where $\gamma_0 = n - tr(\Lambda_0)$, $\gamma_a = n - \mathrm{trace}(\Lambda_a)$, $\nu_1 = \mathrm{trace}(\Lambda_0 - \Lambda_a)$, $\nu_2 = \mathrm{trace}\{(\Lambda_0 - \Lambda_a)^2\}$, $\delta_1 = \mathrm{trace}(\Lambda_a)$, $\delta_2 = \mathrm{trace}(\Lambda_a^2)$, *and* $F_{a,b}$ *is an F-distribution with degrees of freedom* a, b.

The proof is given in the appendix. This result is consistent with the two-moment correction approximate F-test in Section 3.9 of Hastie and Tibshirani (1990).

Remark 1 The degrees of freedom of the above F-test are not the same as those of the regular F-test, e.g. $(\gamma_a - \gamma_0)$ and $(n - \gamma_a)$. The reason is that the smoothing matrix of the penalised splines in a general setting is not a projection matrix. Hence the corresponding degrees of freedom

need to be adjusted. It is easy to see that when the penalty is fixed at 0, the smoothing matrix is idempotent, and we have $\nu_1^2/\nu_2 = \nu_1 = \gamma_a - \gamma_0$, and $\delta_1^2/\delta_2 = \delta_1 = n - \gamma_a$, the degrees of freedom of the conventional F-test derived in Hastie and Tibshirani (1990).

Remark 2 The LRT and restricted likelihood ratio test (RLRT) have been investigated in Crainiceanu and Ruppert (2003), Crainiceanu, Ruppert, and Vogelsang (2003) and Claeskens (2004) for comparing linear models and linear mixed models with one/two variance components. Those proposed (R)LRT statistics based on penalised splines follow asymptotic mixture distributions of independent Chi-square components including χ_0^2, a point mass at 0. Those mixtures are similar to Case 5 in Self and Liang (1987), in which there is one null parameter on boundary, the so-called nonstandard condition. Although asymptotically (R)LRT statistic for the hypotheses is very possible to follow a similar mixture distribution as in those papers above, our proposed $\mathcal{T}_n$ does not follow exact mixtures since a direct mapping does not exist between (R)LRT and $\mathcal{T}_n$. In addition, the statistic $\mathcal{T}_n$ here is to compare a partially linear model (not parametric linear) with a linear mixed model, which includes one random effect in the null models.

2.3. *Bootstrap alternative*

In the previous section, we established that $\mathcal{T}_n$ is asymptotically Chi-square distributed. However, it is difficult if not impossible to find the exact distributions of $\mathcal{T}_n$ for a finite sample size. It is often the case that when the sample size is not sufficiently large, the number of nonparametric terms is not small, and the data are not normally distributed; the asymptotic distribution provides a poor approximation of the true underlying distributions. Here we propose a bootstrap procedure to obtain the critical values for a sample of small size. Specifically,

Step 0 Given original data $\{X_i, Y_i, Z_i, i = 1, \ldots, n\}$,
- (i) Nonparametrically estimate the alternative model (3) and obtain $Z_{a,i}$ and the residuals $\hat{\varepsilon}_{a,i} = Z_i - Z_{a,i}$;
- (ii) Nonparametrically estimate the null model (4) and obtain $Z_{0,i}$;
- (iii) Calculate the test statistic $\mathcal{T}_n$.

Step 1 (i) Generate n iid random variates $\varepsilon_1^*, \ldots, \varepsilon_n^*$ with mean zero, variance 1, and satisfying $|\varepsilon_i^*| \leq C$ for some constant C.
- (ii) Generate the bootstrap sample $Z_i^* = Z_{0,i} + \hat{\varepsilon}_{a,i}\varepsilon_i^*$ for $i = 1, \ldots, n$.

Step 2 For the bootstrap sample $\{X_i, Y_i, Z_i^*, i = 1, \ldots, n\}$,
- (i) Nonparametrically estimate the alternative model (3) and obtain $Z_{a,i}^*$;
- (ii) Nonparametrically estimate the null model (4) and obtain $Z_{0,i}^*$;
- (iii) Calculate the statistic $\mathcal{T}_n^*$.

Step 3 Repeat steps 1 and 2 a total of B times. Compute the empirical distribution of the test statistics based on the bootstrap samples $\{\mathcal{T}_{1n}^*, \ldots, \mathcal{T}_{Bn}^*\}$.

Step 4 Let $\hat{t}_\xi^W$ be the $(1 - \xi)$ quantile of the bootstrap empirical distribution. Reject the null hypothesis (model (4)) in favour of the alternative hypothesis (model (3)) at ξ level if $\mathcal{T}_n > \hat{t}_\xi^W$.

The above approach is based on the idea of a wild bootstrap, originally proposed by Wu (1986) for linear regression and incorporated to nonparametric setups by Härdle and Mammen (1993). In our implementation, we use a two-point distribution $\gamma\delta_a + (1 - \gamma)\delta_b$ for generating ε_i^*, where $\gamma = (5 + \sqrt{5})/10$, $a = (1 - \sqrt{5})/2$, $b = (1 + \sqrt{5})/2$, and δ_a denotes the point measure at a. See Härdle and Marron (1991) for a detailed discussion on the wild bootstrap.

3. General additive models

In this section, we consider the problem of determining linear components in a general additive model with multiple components

$$Z = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \varepsilon, \tag{8}$$

where X_j $(j = 1, \ldots, p)$ $(p \geq 3)$ are univariate variables in a compact support interval that we assume to be [0, 1] without loss of generality. We are interested in whether some of the functions are linear and consider two basic problems under this setting.

First, we consider the simpler problem of testing a given specific subset of the functions being linear. For notational convenience, suppose we want to simultaneously test $f_1(\cdot), \ldots, f_k(\cdot)$ to be linear, where $k < p$. The procedure is almost exactly the same as the two regressor case presented in the previous section. We first fit the model under the null hypothesis

$$Z = c + \beta_1 X_1 + \cdots + \beta_k X_k + f_{k+1}(X_{k+1}) + \cdots + f_p(X_p) + \varepsilon,$$

and the general model (8) under the alternative hypothesis using the penalised spline method, then obtain their corresponding fitted values $Z_{0,i}$ and $Z_{a,i}$. The test statistic $\mathcal{T}_n$ given in Equation (5) can be used. With similarly constructed A_0 and A_a matrices, Theorem 1 continues to hold in this case. As in Section 2.3, finite sample critical values of the test statistic can be obtained through a bootstrap method.

More difficult is the question of determining the subset of functions that are linear without prior indication. Here we propose a step-wise procedure similar to the variable selection procedure in linear regression. Specifically, we start with full nonparametric additive model (8) and iterate the following steps:

Backward determination procedure for $k = 1, 2, \ldots$

(1) Suppose at step k we have identified $k - 1$ linear components in the model. Without loss of generality, let the first $k - 1$ functions be linear.
(2) For $j = k, \ldots, p$, compute the test statistic $\mathcal{T}_{n,j}$ in Equation (5) for comparing the null model

$$\begin{aligned} Z = c + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + f_k(X_k) + \cdots + f_{j-1}(X_{j-1}) \\ + \beta_j X_j + f_{j+1}(X_{j+1}) + \cdots + f_p(X_p) + \varepsilon, \end{aligned}$$

and the alternative model

$$Z = c + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + f_k(X_k) + \cdots + f_p(X_p) + \varepsilon,$$

and obtain their corresponding p-values. If the minimum of the $p - k$ p-value is less than a pre-determined level α^* (e.g. 0.01), we identify the corresponding component as a linear component and relabel the variable as X_k and repeat this step. Stop if none of the p-values are less than α^*.

A similar forward selection method can be used by starting with a full linear model and identifying the nonlinear components one variable at a time.

4. Simulation

We conduct simulation experiments in this section to explore the numerical performance of the proposed procedure.

Example 4.1 In this example, observations are generated from a two-regressor model

$$Z_i = f(X_i) + g(Y_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n.$$

We use $X_i \sim Unif[0, 1]$ and $Y_i \sim Unif[0, 1]$, which are independent. $\varepsilon_i \sim N(0, \sigma)$ for $\sigma = 0.25, 0.5, 1$. Two sample sizes were considered: $n = 50, 100$. The functions $f(\cdot)$ and $g(\cdot)$ are chosen to be one of the following four pairs: (a) $f(x) = 3\sin(\pi x)$, $g(y) = 2\sin(2\pi y)$; (b) $f(x) = 3x$, $g(y) = 2\sin(2\pi y)$; (c) $f(x) = 3x^3 + (x - 1/4)(x - 1/2)(x - 3/4)$, $g(y) = 2\sin(2\pi y)$; and (d) $f(x) = 4x$, $g(y) = 3y$. We are interested in testing the linearity of the function $f(x)$.

In this experiment, $K_x = 17$ and $K_y = 15$ knots are used to fit models (3) and (4). At each configuration and sample size, 500 independent sets of data were generated. For the bootstrap procedure, $B = 200$ bootstrap samples were used for each data set.

The simulation results are summarised in Table 1. Note that in cases (a) and (c), the alternative hypothesis is true, and the corresponding values in Table 1 indicate that the null hypothesis is almost always rejected regardless of the sample size and the variances of the data. In cases (b) and (d), the null hypothesis is true, and the corresponding values in Table 1 indicate that these empirical levels are close to the nominal levels. It is expected that the levels are closer to the nominal levels with larger sample sizes and small variances.

In studying the power of the test, we use $f(x, c) = 3x + c(x - 1/4)(x - 1/2)(x - 3/4)$ and $g(y) = 2\sin(\pi y)$ with $c = 0, 5, 10, 20, 25$. Panel (a) of Figure 1 depicts the function $f(x, c)$ for $c = 5, 10, 20, 25$. We test the linearity of the function $f(x, c)$ and see the effects of c on the percentage of rejections of the null hypothesis. The nominal level of the test is 0.05 in each case. Panels (b), (c), and (d) of Figure 1 present the power of the three different tests for different variance configurations. The solid, dashed, and dotted lines represent the power of the test as a function of c when the variance of the error is 0.25^2, 0.5^2, and 1. The power of the test increases as the function becomes more and more nonlinear (i.e. c increases), as one would expect. It is also seen that the variance of the noise plays a very important role for the power behaviour. More importantly, it is seen that the bootstrap method has a smaller power than the ones based on asymptotic distributions. However, it may not be viewed as a setback for the bootstrap method,

Table 1. Percentage of times that the null hypothesis was rejected using the asymptotic Chi-square test (Chi), the asymptotic F-test (F), and the bootstrap test (B).

		(a)			(b)			(c)			(d)		
n	σ	Chi	F	B	Chi	F	B	Chi	F	B	Chi	F	B
						$\alpha = 0.05$							
50	0.25	100	100	100	3.3	3.3	0	100	100	100	0	0	0
	0.5	100	100	100	6.4	7.5	7.5	100	97.9	100	5.7	8	9.1
	1	100	100	100	5	7	7	84	81	82	7	8	9
100	0.25	100	100	100	6.2	8.4	5.2	100	100	100	7.4	8.8	5.3
	0.5	100	100	100	2	6	2	100	100	100	5	7	5
	1	100	100	100	8.1	10.1	7.1	98	93.9	97	3	6.1	4
						$\alpha = 0.1$							
50	0.25	100	100	100	11.7	13.7	13.9	100	100	100	0	0	0
	0.5	100	100	100	11	19.1	13.8	100	100	100	7.6	8.7	8.7
	1	100	100	100	11	11	10	90	85	88	12	12	12
100	0.25	100	100	100	9.4	8.6	16.3	100	100	100	11.3	14.4	10.3
	0.5	100	100	100	13	15	15	100	100	100	8	11	5
	1	100	100	100	11	18	12	100	98	99	11	18	14

Note: Five hundred independent sets of data and 200 bootstrap samples were generated from each data set.

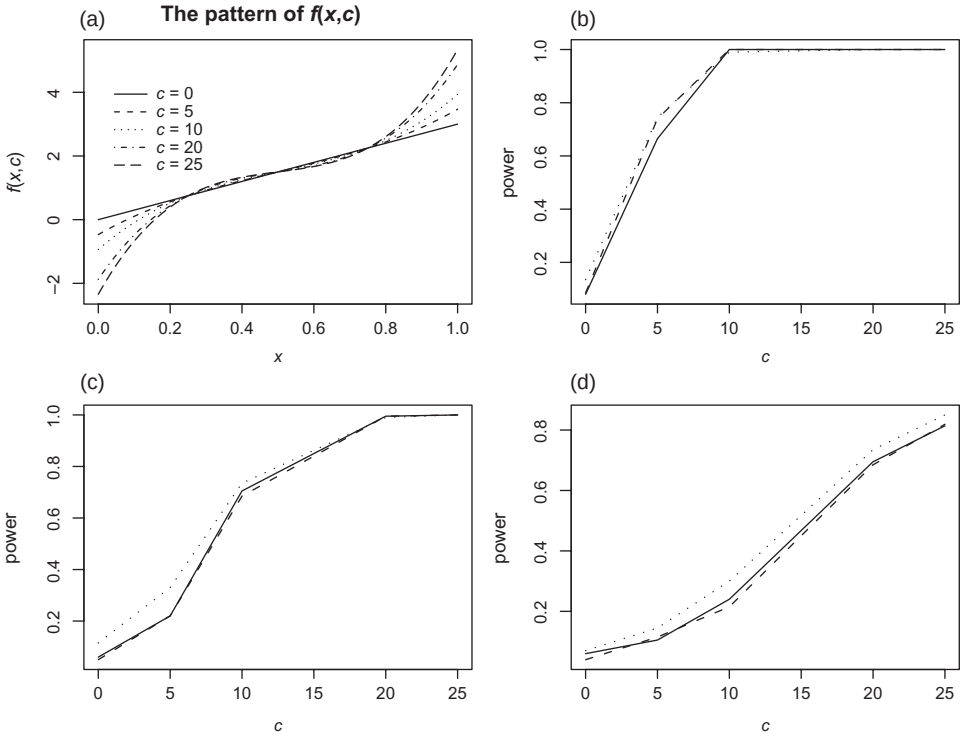

Figure 1. (a) The pattern of $f(x, c)$ for different values c. The power of the tests for different cases; (b) $\sigma = 0.25$, (c) $\sigma = 0.5$, and (d) $\sigma = 1$. The solid, broken, and dotted lines represent the results corresponding to the bootstrap test, the asymptotic Chi-square test, and the asymptotic F-test.

as Table 1 under (b) shows that the tests based on the asymptotic distributions tend to have higher critical values and hence tend to be rejected more often.

Example 4.2 In this example, we generate data from the four-component model

$$Z_i = f_1(X_{1i}) + f_2(X_{2i}) + f_3(X_{3i}) + f_4(X_{4i}) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where X_{1i}, X_{2i}, X_{3i}, and X_{4i} are independent uniform [0, 1] random variates and $\varepsilon_i \sim N(0, \sigma^2)$. The four functions are set to the following three combinations:

(e) $f_1(x_1) = 3\sin(\pi x_1)$, $f_2(x_2) = 2\sin(2\pi x_2)$, $f_3(x_3) = 0.45\sin(\pi x_3)$, $f_4(x_4) = 3x_4^3 + (x_4 - 1/4)(x_4 - 1/2)(x_4 - 3/4)$;
(f) $f_1^*(x_1) = 3x_1$, $f_2(x_2) = 2\sin(2\pi x_2)$, $f_3(x_3) = 0.45\sin(\pi x_3)$, $f_4(x_4) = 3x_4^3 + (x_4 - 1/4)(x_4 - 1/2)(x_4 - 3/4)$;
(g) $f_1^*(x_1) = 4x_1$, $f_2^*(x_2) = 3x_2$, $f_3(x_3) = 0.45\sin(\pi x_3)$, $f_4(x_4) = 3x_4^3 + (x_4 - 1/4)(x_4 - 1/2)(x_4 - 3/4)$.

We are interested in checking the linearity of $f_1(\cdot)$ and $f_2(\cdot)$. The cases of $\sigma = 0.25$, 0.5, and 1, and nominal levels 0.05 and 0.1 are considered.

We, therefore, first check the linearity of $f_1(\cdot)$. The simulation results are presented in Table 2, which shows a similar pattern as in Table 1, though the empirical levels are not as good as those in the two-component case. It seems that the extra terms $f_3(x_3)$ and $f_4(x_4)$ may play a negative role, as they introduce extra variability in the data, hence making the linearity of $f_1(X_1)$ less

Table 2. Percentage of times that the null hypothesis, linear function $f_1(x_1)$, was rejected using the Chi-squared approximation (Chi), F-approximation (F), and bootstrap (B) for Example 4.2.

		(e)			(f)			(g)		
n	σ^2	Chi	F	B	Chi	F	B	Chi	F	B
					$\alpha = 0.05$					
100	0.25	100	100	100	2.6	8.3	4	100	100	100
	0.5	100	100	100	3.6	9.1	3.6	100	100	100
	1	100	100	100	4.4	7.1	6.7	98.8	97.7	98.3
200	0.25	100	100	100	5.9	8.3	7.7	100	100	100
	0.5	100	100	100	2.7	6.7	5.9	100	100	100
	1	100	100	100	3.2	8.8	3.8	100	99.4	100
					$\alpha = 0.1$					
100	0.25	100	100	100	9.2	12.1	11.8	100	100	100
	0.5	100	100	100	6	15.9	12.6	100	100	100
	1	100	100	100	3.6	13.7	8.9	98.1	96.2	98.7
200	0.25	100	100	100	6	9.9	13.2	100	100	100
	0.5	100	100	100	2.3	13.8	8.7	100	100	100
	1	100	100	100	1.6	13.3	11.7	100	100	100

Note: Five hundred independent sets of data and 200 bootstrap samples were generated from each data set.

Table 3. Percentage of times that the null hypothesis that $f_2(x)$ is linear is rejected using the Chi-squared approximation (Chi), F-approximation (F), and bootstrap (B) for Example 4.2.

		(f1)			(g1)		
n	σ^2	Chi	F	B	Chi	F	B
				$\alpha = 0.05$			
100	0.25	100	100	100	3.5	6.5	6.8
	0.5	100	95	100	4.6	5	7
	1	100	90	100	5.7	6.7	10.5
200	0.25	100	100	100	4.6	5	8
	0.5	100	100	100	6.8	8.7	9
	1	100	100	100	7	6.5	7.5
				$\alpha = 0.1$			
100	0.25	100	100	100	9.5	11.5	12
	0.5	100	100	100	10.5	10	12.5
	1	100	100	100	12.5	13.2	15
200	0.25	100	100	100	10	11.5	13.5
	0.5	100	100	100	10	13.5	12.5
	1	100	100	100	12.5	13.0	12.5

Note: Five hundred independent sets of data and 200 bootstrap samples were generated from each data set.

detectable. The power of the test in four components is similar to that in two components (not shown here). We then further check the linearity of $f_2(\cdot)$ for cases (f) and (g) by using the procedure proposed in Section 3. The simulation results are presented in Table 3 under columns (f1) and (g1). These empirical results indicate that the null hypothesis, linear function $f_2(x_2)$, for case (f) was rejected with remarkably high power, while case (g) was accepted with empirical levels close to but mostly slightly higher than the nominal level. Overall, the asymptotic Chi-square test gives the most favourable results.

We now run a simulation to study the performance of the stepwise determination procedure as outlined in Section 3. The simulation results are presented in Table 4, for which we applied the asymptotic Chi-square test for 1000 independent sets of each setting. It is seen that the

Table 4. Times of each component being classified as a linear one using the Chi-squared approximation for Example 4.2.

		(e)				(f)				(g)			
n	σ^2	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
						$\alpha = 0.05$							
100	0.2	0	0	1	0	965	0	1	0	951	950	0	0
	0.5	0	0	287	0	966	0	210	0	962	946	156	0
	1	0	0	775	37	970	0	702	28	973	962	682	40
200	0.2	0	0	0	0	970	0	0	0	971	964	0	0
	0.5	0	0	63	0	979	0	26	0	981	970	13	0
	1	0	0	637	0	974	0	607	0	978	982	494	0
						$\alpha = 0.1$							
100	0.2	0	0	1	0	949	0	0	0	918	905	0	0
	0.5	0	0	200	0	930	0	158	0	938	943	98	0
	1	0	0	728	9	951	0	655	17	942	947	593	32
200	0.2	0	0	0	0	966	0	0	0	951	945	0	0
	0.5	0	0	39	0	969	0	12	0	967	960	10	0
	1	0	0	569	0	974	0	494	0	955	970	431	0

Note: Thousand independent sets of data were generated from each setting.

proposed procedure can always powerfully identify a linear pattern when the component is truly linear. On the other hand, we are also interested in how the proposed procedure erroneously classifies a nonlinear component as a linear one. When checking at the four nonlinear functions $f_1(\cdot), \ldots, f_4(\cdot)$, it is clear that $f_3(\cdot)$ is the easiest one to be detected as linear. The simulation results confirmed this observation. The possibility of this mistreatment is higher when the variance of the model errors gets larger or when more nonlinear components are involved. This may not be surprising because with larger noise, it is more difficult to detect the pattern of the function. Overall, these simulation results indicate that the proposed procedure is very promising in determining the subset of linear functions.

5. A real-data example

Here we use the proposed approach to analyse a data set from an environmental study (Hauser, Meeker, Duty, Silva, and Calafat 2006) conducted at Massachusetts General Hospital (MGH), to study the relationships between semen quality and chemical exposure (naphthol) and several covariates, including abstinence time, age, body mass index (BMI), and race (white or non-white). Naphthol is a metabolite of both carbaryl (a broad-spectrum insecticide) and naphthalene formed from incomplete combustion of hydrocarbons from fuel and other sources.

Liang, Thurston, Ruppert, Apanasovich, and Hauser (2008) proposed an additive model to study the relationship between semen quality and other risk factors in another environmental study, where they treated age and BMI as nonlinear predictors. As an illustration, we consider the following model:

$$\log(\text{sperm concentration}) = f_1(\text{age}) + f_2(\text{lnsgln}) + f_3(\text{BMI}) + \alpha_4 \text{abstin} + \alpha_5 \text{race}.$$

Here, lnsgln denotes the log-transformed 1-naphthol adjusted for specific gravity, and abstin denotes abstinence time. We are concerned with the linearity of f_1, f_2, and f_3. We used the proposed test to check the linearity of f_1, took $K_{z1} = 15$ and $K_{z2} = 18$, and generated 500 samples in a bootstrap resample approach. The results from both the asymptotic Chi-square test

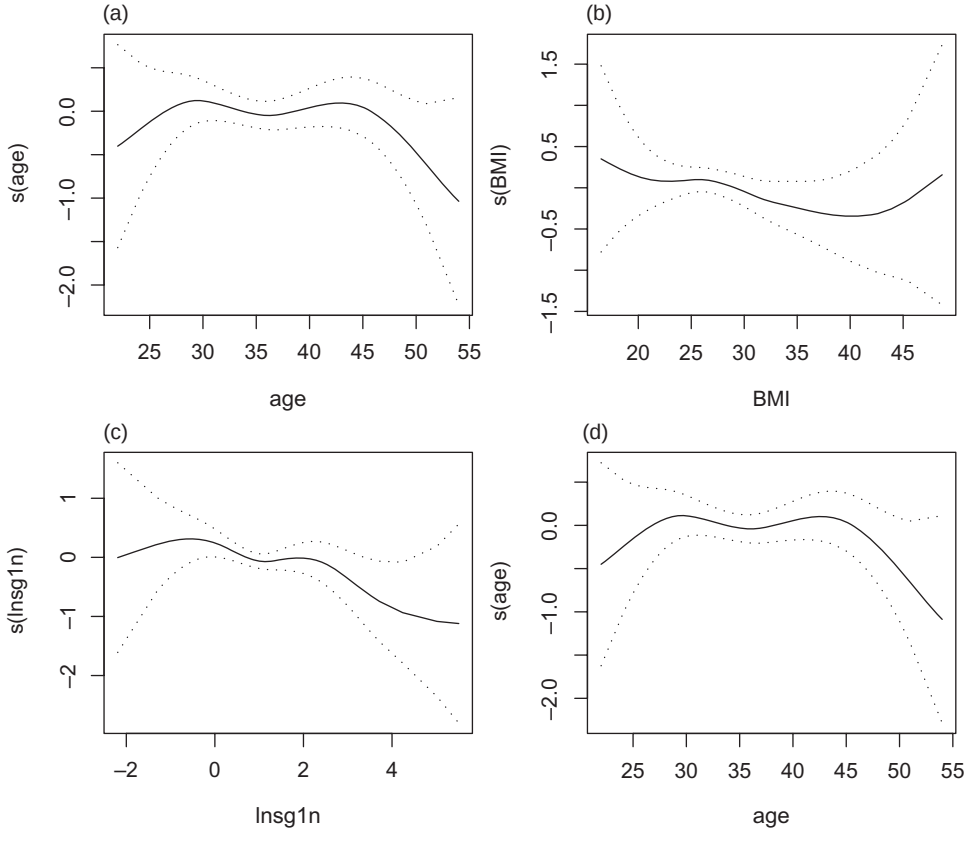

Figure 2. The patterns of age, BMI, and lnsgln with ±s.e. assuming that all three components are nonlinear (a)–(c), or only the age component is nonlinear (d) using the R gam function for the semen study.

and the asymptotic F-test indicate that we should reject the linear assumption on f_1. The bootstrap p-value is $<10^{-4}$. Furthermore, we check the linearity of f_2 and f_3 and repeat the imputation except for using *age* as a nonlinear component. The bootstrap p-values are 0.294 and 0.371, and the results of the asymptotic Chi-square and F-tests indicate a linearity of f_2 and f_3. The stepwise determination procedure proposed in Section 3 is used to determine the linear components. With $\alpha = 0.05$ the procedure identified that the function associated with the age variable is nonlinear, while all other functions are linear.

Hence both approaches indicate that log(sperm concentration) depends nonlinearly on age, but linearly on lnsgln and BMI. We present the functions associated with age, BMI, and lnsgln using a model assuming all three components are nonlinear in Figure 2(a)–(c) and a model with only the age component is nonlinear in Figure 2(d). The patterns of (a) and (d) are almost identical. These plots also confirm the finding that age is the only nonlinear component in the model.

6. Discussion

We proposed a testing procedure to determine the linear components in additive models. The basic test is based on the sum squared differences of the fitted values under the null hypothesis that one set of components are linear functions in an additive model and the alternative hypothesis that one of the components in the set is not linear. We used the penalised spline to estimate the

nonparametric components and therefore avoided iterative procedures. An approximation of the asymptotic distribution of the test statistics under null hypothesis is derived and a bootstrap version of the test statistic was also proposed. In addition, a forward selection procedure is proposed to determine all linear components in an additive model. Simulation experiments and real-data analysis show strong evidence that our procedure is accurate and useful in real applications.

Acknowledgements

Chen's research was partially supported by NSF grants DMS-0905763, DMS-0915139, and DMS-0800183. Liang's research was partially supported by NSF grants DMS-0806097 and DMS-1007167. Wang's research was partially supported by the University of Illinois at Chicago WISER Fund Award and NSF ADVANCED IT Award (SBE-0546843).

References

Aerts, M., Claeskens, G., and Wand, M.P. (2002), 'Some Theory for Penalized Spline Generalized Additive Models', *Journal of Statistical Planning and Inference*, 103, 455–470.

Aitken, A.C. (1950), 'On the Statistical Independence of Quadratic Forms in Normal Variates', *Biometrika*, 37, 93–96.

Barry, D. (1993), 'Testing for Additivity of a Regression Function', *The Annals of Statistics*, 21, 235–254.

de Boor, C. (2001). *A Practical Guide to Splines*, New York: Springer-Verlag.

Box, G.E. (1954), 'Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification', *The Annals of Mathematical Statistics*, 25, 290–302.

Brumback, B.A., Ruppert, D., and Wand, M. (1999), 'Comment on "Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-based Prior" by Shively, Kohn and Wood', *Journal of the American Statistical Association*, 94, 794–797.

Cai, Z.W., Fan, J.Q., and Li, R.Z. (2000), 'Efficient Estimation and Inferences for Varying-Coefficient Models', *Journal of the American Statistical Association*, 95, 888–902.

Chen, R., Liu, J.S., and Tsay, R.S. (1995), 'Additivity Tests for Nonlinear Autoregressive Models', *Biometrika*, 82, 369–383.

Claeskens, G. (2004), 'Restricted Likelihood Ratio Lack-of-Fit Tests Using Mixed Spline Models', *Journal of the Royal Statistical Society, Series B*, 66, 909–926.

Claeskens, G., Krivobokova, T., and Opsomer, J. (2008), 'Asymptotic Properties of Penalized Spline Estimator', *Biometrika*, 96, 529–544.

Crainiceanu, C.M., and Ruppert, D. (2003), 'Likelihood Ratio Tests in Linear Mixed Models with One Variance Component', *Journal of the Royal Statistical Society, Series B*, 66, 165–185.

Crainiceanu, C.M., Ruppert, D., and Vogelsang, T.J. (2003), 'Some Properties of Likelihood Ratio Tests in Linear Mixed Models', manuscript, www.orie.cornell.edu/∼davidr/papers.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M.P. (2005), 'Exact Likelihood Ratio Test for Penalized Splines', *Biometrika*, 92, 91–103.

Cushman, M., Kuller, L.H., Prentice, R., Rodabough, R.J., Psaty, B.M., Stafford, R.S., Sidney, S., Rosendaal, R.F., and for the Women's Health Initiative Investigators. (2004), 'Estrogen Plus Progestin and Risk of Venous Thrombosis', *Journal of the American Medical Association*, 292, 1573–1580.

Dominici, F., McDermott, A., Zeger, S.L., and Samet, J.M. (2002), 'On the Use of Generalized Additive Models in Time-Series of Air Pollution and Health', *American Journal of Epidemiology*, 156, 193–203.

Eilers, P.H.C., and Marx, B.D. (1996), 'Flexible Smoothing with B-Splines and Penalties (with comments)', *Statistical Science*, 11, 89–121.

Eubank, R.L., and Speigelman, C. (1990), 'Testing the Goodness-of-Fit Linear Models Via Nonparametric Regression Techniques', *Journal of the American Statistical Association*, 85, 387–392.

Hall, P., and Opsomer, J.D. (2005), 'Theory for Penalised Spline Regression', *Biometrika*, 92, 105–118.

Härdle, W., and Mammen, E. (1993), 'Comparing Nonparametric Versus Parametric Regression Fits', *The Annals of Statistics*, 21, 1926–1947.

Härdle, W., and Marron, S. (1991), 'Bootstrap Simultaneous Error Rates for Nonparametric Regression', *The Annals of Statistics*, 19, 778–796.

Härdle, W., and Tsybakov, A.B. (1995), 'Additive Nonparametric Regression on Principal Components', *Journal of Nonparametric Statistics*, 5, 157–184.

Härdle, W., Liang, H., and Gao, J.T. (2000), *Partially Linear Models*, Heidelberg: Springer Physica-Verlag.

Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004), 'Bootstrap Inference in Semiparametric Generalized Additive Models', *Econometric Theory*, 20, 265–300.

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

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

Hastie, T.J., and Tibshirani, R.J. (1993), 'Varying-Coefficient Models (with discussion)', *Journal of the Royal Statistical Society, Series B*, 55, 757–796.

Hauser, R., Meeker, J.D., Duty, S., Silva, M.J., and Calafat, A.M. (2006), 'Altered Semen Quality in Relation to Urinary Concentrations of Phthalate Monoester and Oxidative Metabolites', *Epidemiology*, 17, 682–691.
Hjellvik, V., and Tjøstheim, D. (1996), 'Nonparametric Statistics for Testing of Linearity and Serial Independence', *Journal of Nonparametric Statistics*, 6, 223–251.
Hsu, C.C., Kao, L.W.H., Coresh, J., Pankow, J.S., Marsh-Manzi, J., Boerwinkle, E., and Bray, M.S. (2005), 'Apolipoprotein E and Progression of Chronic Kidney Disease', *Journal of the American Medical Association*, 93, 2892–2899.
Huang, J.Z. (2003), 'Local Asymptotics for Polynomial Spline Regression', *The Annals of Statistics*, 31, 1600–1625.
Huang, L.S., and Chen, J. (2008), 'Analysis of Variance, Coefficient of Determination, and F-test for Local Polynomial Regression', *The Annals of Statistics*, 36, 2085–2109.
Ichimura, H. (1993), 'Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models', *Journal of Econometrics*, 58, 71–120.
Li, Q., and Rachine, J.S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton: Princeton University Press.
Li, Y., and Ruppert, D. (2008), 'On the Asymptotics of Penalized Splines,' *Biometrika*, 95, 415–436.
Liang, H., Thurston, S., Ruppert, D., Apanasovich, T., and Hauser, R. (2008), 'Additive Partial Linear Models with Measurement Errors', *Biometrika*, 95, 667–678.
Linton, O.B., and Nielsen, J.P. (1995), 'A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration,' *Biometrika*, 82, 93–101.
Liu, R., and Yang, L. (2010), 'Spline-Backfitted Kernel Smoothing of Additive Coefficient Model', *Econometric Theory*, 26, 29–59.
Ma, S., and Yang, L. (2011), 'Spline-Backfitted Kernel Smoothing of Partially Linear Additive Model', *Journal of Statistical Planning and Inference*, 141, 204–219.
Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, New York: Academic Press.
Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008), 'Non-parametric Small Area Estimation Using Penalized Spline Regression', *Journal of the Royal Statistical Society, Series B*, 70, 265–286.
Ruppert, D. (2002), 'Selecting the Number of Knots for Penalized Splines', *Journal Computational and Graphical Statistics*, 11, 735–757.
Ruppert, D., and Carroll, R. (2000), 'Spatially-Adaptive Penalties for Spline Fitting,' *Australian Journal of Statistics*, 42, 205–223.
Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*, New York: Cambridge University Press.
Self, S., and Liang, K. (1987), 'Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions', *Journal of the American Statistical Association*, 82, 605–610.
Song, Q., and Yang, L. (2010), 'Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Models with Simultaneous Confidence Band', *Journal of Multivariate Analysis*, 101, 2008–2025.
Sperlich, S., Tjøstheim, D., and Yang, L. (2002), 'Nonparametric Estimation and Testing of Interaction in Additive Models', *Econometric Theory*, 18, 197–251.
Stone, C.J. (1986), 'The Dimensionality Reduction Principle for Generalized Additive Models', *The Annals of Statistics* 14, 590–606.
Wahba, G. (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, 59. Society for Industrial and Applied Mathematics (SIAM).
Wand, M.P. (1999), 'On the Optimal Amount of Smoothing in Penalized Spline Regression', *Biometrika*, 86, 936–940.
Wang, L., and Yang, L. (2007), 'Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model', *The Annals of Statistics*, 35, 2474–2503.
Wu, C.F.J. (1986), 'Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis', *The Annals of Statistics*, 14, 1261–1343.
Xia, Y., Zhang, W., and Tong, H. (2004), 'Efficient Estimation for Semivarying-Coefficient Models,' *Biometrika*, 91, 661–681.
Yu, Y., and Ruppert, D. (2002), Penalized Spline Estimation for Partially Linear Single Index Models', *Journal of the American Statistical Association*, 97, 1042–1054.

Appendix

For real vectors $\mathbf{x}, \mathbf{y} \in R^n$, define the inner product and L_2 norm as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, $\|\mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x}$. Denote the number of truncated power basis functions for the alternative model and null model by $K_a = 5 + K_x + K_y$, $K_0 = 4 + K_y$, respectively.

We need the following assumptions.

(A1) The two component functions $(\cdot)$, $g(\cdot) \in C^{(3)}[0, 1]$.
(A2) The densities of X and Y are bounded above and below on $[0, 1]$.
(A3) The numbers of knots satisfy $K_x, K_y \sim n^{1/6}$.

Writing the additive model as $\mathbf{Z} = f(\mathbf{x}) + g(\mathbf{y}) + \boldsymbol{\varepsilon} = \mathbf{m} + \boldsymbol{\varepsilon}$, we have $\hat{\mathbf{Z}}_a - \hat{\mathbf{Z}}_0 = (A_a - A_0)(\mathbf{m} + \boldsymbol{\varepsilon}) = (\mathbf{m}_a - \mathbf{m}) - (\mathbf{m}_0 - \mathbf{m}) + (A_a - A_0)\boldsymbol{\varepsilon}$, where $\mathbf{m}_a = \boldsymbol{X\beta} + \boldsymbol{Y\gamma} + \boldsymbol{\Xi}(x)\boldsymbol{b} + \boldsymbol{\Xi}(y)\boldsymbol{d}$ and $\mathbf{m}_0 = \boldsymbol{X}_{[1,2]}\boldsymbol{\beta}_{[1,2]} + \boldsymbol{Y\gamma} + \boldsymbol{\Xi}(y)\boldsymbol{d}$. Similar to (Ruppert et al. 2003, Section 6.4), it can be shown that $E(\hat{\mathbf{Z}}_a)A_a\mathbf{m} = \mathbf{m}_a$ and $E(\hat{\mathbf{Z}}_0) = A_0\mathbf{m} = \mathbf{m}_0$.

If the two bias squares, $\|\mathbf{m}_a - \mathbf{m}\|^2$ and $\|\mathbf{m}_0 - \mathbf{m}\|2$, have negligible order compared with $\|(A_a - A_0)\boldsymbol{\varepsilon}\|^2$, then the asymptotical distribution of $\|(A_a - A_0)\mathbf{Z}\|^2$ is the same as $\|(A_a - A_0)\boldsymbol{\varepsilon}\|^2$. Here, we only show the results for bias $\mathbf{m}_a - \mathbf{m}$ and the derivation for the order of $\mathbf{m}_0 - \mathbf{m}$ can be obtained similarly.

LEMMA A.1 *Under Assumptions* (A1)–(A3), *we have* $\|\mathbf{m}_a - \mathbf{m}\|^2 = O_p(K_a^{-6} + n^{-1})$.

Proof First, we start with the smoothing matrix $A_a = C_a(C_a^{\mathrm{T}}C_a + B_a)^{-1}C_a^{\mathrm{T}}$. In the following, we assume that, as n goes to infinity, the limit of the matrix $A_n^* = \frac{1}{n}C_a^{\mathrm{T}}C_a$ is a positive definite matrix M_L with bounded eigenvalues in probability. This is a standard assumption for penalised splines that has been used in Appendix B of Yu and Ruppert (2002). Under this assumption, all eigenvalues $\{\lambda_{j,L}\}_{j=1}^{K_a}$ of M_L satisfy that for any $1 \le j \le K_a$, $\lambda_{j,L}/n$ is less than 1 when n is sufficiently large. This is also a required condition (A4) in Claeskens et al. (2008). In practice, A_n^* may not be positive definite, but one can add very small positive constants to the diagonal elements to make it positive definite, see Ruppert (2002). In the following matrix, A_n^* will be treated as a nonsingular matrix.

Because the diagonal matrix B_a has constant diagonals, the restriction of A_n^* implied that all eigenvalues of $(1/n)A_n^{*-1}B_a$ are less than 1 when n is large. Based on the standard matrix theory, the following expansion is valid: $(I + (1/n)A_n^{*-1}B_a)^{-1} = I - (1/n)A_n^{*-1}B_a + o_p((1/n)A_n^{*-1}B_a)$. Then A_a can be expanded as follows:

$$\begin{aligned} A_a &= C_a\left(I + \frac{1}{n}A_n^{*-1}B_a\right)^{-1}(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}} \\ &= C_a\left\{I - \frac{1}{n}A_n^{*-1}B_a + o_p\left(\frac{1}{n}A_n^{*-1}B_a\right)\right\}(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}. \end{aligned}$$

Denoting $P_a = C_a(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}$, $G^* = C_a(A_n^{*-1}B_a)(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}$, then the above equation can be simplified as

$$A_a = P_a - \frac{1}{n}G^* + o_p\left(\frac{1^*}{n}\right). \tag{A1}$$

Correspondingly, the asymptotic behaviour of the bias term A_n can be obtained through its major terms, i.e. $\mathbf{m}_a - \mathbf{m} = (A_a - I)\mathbf{m} \sim (P_a - I)\mathbf{m} - (1/n)G^*\mathbf{m}$ in which $a_n \sim b_n$ means that $a_n/b_n = O_p(1)$. Taking a squared L_2 norm on both sides, we have

$$\|(A_a - I)\mathbf{m}\|^2 \sim \|(P_a - I)\mathbf{m}\|2 - \frac{2}{n}\langle(P_a - I)\mathbf{m}, G^*\mathbf{m}\rangle + n^{-2}\|G^*\mathbf{m}\|^2. \tag{A2}$$

The first term on the right-hand side is the squared bias of using regression splines to approximate the smooth function $\mathbf{m}$. The bias order for univariate regression splines depends on the knots number and spline order, similar to the B-spline approximation in Theorem 6 on Page 149 of de Boor (2001) and Huang (2003, Theorem 5.1). If the knots are equally spaced, the B-spline basis is equivalent to the truncated power basis. For the given knots order in Assumption (A3) and quadratic degree (order = 3), we have $\sup_{x\in[0,1]}\|(P_a - I)\mathbf{m}\| = O_p(K_a^{-3})$. The second term on the right-hand side of Equation (A2) is dominated by the other two terms through Cauchy–Schwartz inequality, i.e.

$$\frac{2}{n}\langle(P_a - I)\mathbf{m}, G^*\mathbf{m}\rangle \le 2 \cdot \{n^{-2}\|^*\mathbf{m}\|^2 \cdot \|(P_a - I)\mathbf{m}\|^2\}^{1/2}.$$

The third term on the right-hand side of Equation (A2) is $n^{-2}\|G^*\mathbf{m}\|^2 = n^{-2}(G^*\mathbf{m})^{\mathrm{T}}G^*\mathbf{m} = n^{-2}\mathbf{m}^{\mathrm{T}}G^{*2}\mathbf{m}$. Note that

$$\begin{aligned} G^{*2} &= \{C_a(A_n^{*-1}B_a)(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}\}^{\mathrm{T}}C_a(A_n^{*-1}B_a)(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}} \\ &= C_a(C_a^{\mathrm{T}}C_a)^{-1}(A_n^{*-1}B_a)C_a^{\mathrm{T}} \cdot C_a(A_n^{*-1}B_a)(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}} \\ &= C_aA_n^{*-2}B_a^2(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}. \end{aligned}$$

Suppose the eigen-decomposition of the matrix A_n^* is $A_n^* = \Gamma V_n\Gamma^{\mathrm{T}}$, in which matrix Γ is orthogonal and V_n is diagonal. Hence $A_n^{*-2} = \Gamma V_n^{-2}\Gamma^{\mathrm{T}}$. ■

Because matrix B_a^2 is a diagonal matrix with constant diagonal elements, and V_n^{-2} is diagonal with bounded eigenvalues, neither diagonal matrix affects the asymptotic order of G^{*2}. In other words, G^{*2} has the same asymptotic order as $C_a(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}$, which is the projection matrix P_a defined above. Then $P_a\mathbf{m}$ is the projection of $\mathbf{m}$ onto the truncated power spline space, and $n^{-2}\mathbf{m}^{\mathrm{T}}G^{*2}\mathbf{m} \sim n^{-2}\mathbf{m}^{\mathrm{T}}P_a\mathbf{m} = n^{-2}\mathbf{m}^{\mathrm{T}}P_a^{\mathrm{T}}P_a\mathbf{m} = n^{-2}\|P_a\mathbf{m}\|^2$. Applying the triangle inequality of L_2 norm, we have $\|P_a\mathbf{m}\|^2 \le \|\mathbf{m}\|^2 + \|(P_a - I)\mathbf{m}\|^2$. For a second-order smooth function $\mathbf{m}$ on a compact interval $[0, 1]$, $n^{-2}\|\mathbf{m}\|^2 \sim n^{-1}$, so $n^{-2}\|G^*\mathbf{m}\|^2 = O_p(n^{-1} + n^{-2}K_a^{-6})$. The order of the sum of the three terms in Equation (A2) leads to the bias order as follows $\|(A_a - I)\mathbf{m}\|^2 = O_p(K_a^{-6} + n^{-1})$. We complete the proof.

Proof of Theorem 2.1 Assume that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma_\varepsilon^2 I_n)$, where I_n is the $n \times n$ identity matrix. Denote the symmetric matrix $(A_a - A_0)^2$ as Λ. Based on Cochran's theorem (Mardia, Kent, and Bibby 1979, p. 68), the quadratic form $\boldsymbol{\varepsilon}^{\mathrm{T}}\Lambda\boldsymbol{\varepsilon}$ follows

a weighted Chi-square distribution, i.e. $\boldsymbol{\varepsilon}^{\mathrm{T}}\Lambda\boldsymbol{\varepsilon} \sim \sigma_\varepsilon^2 \sum_{i=1}^{K'} \lambda_i \cdot \chi^2(1)$, where $\lambda_i' s$ are the eigenvalues of the matrix Λ and K' is the rank of Λ.

By Box (1954, Theorem 3.1), the above weighted Chi-square can be approximated by a Chi-square distribution of $\mathbf{c}\chi^2(b)$ with parameters $\mathbf{c}$, b defined in Equation (6). The two parameters can be expressed as the ratio of traces

$$\mathbf{c} = \sigma_\varepsilon^2 \frac{\text{trace}(\Lambda^2)}{\text{trace}(\Lambda)}, \quad b = \frac{\text{trace}^2(\Lambda)}{\text{trace}(\Lambda^2)}. \tag{A3}$$

Lemma A.1 shows that the bias is of order $o_p(1)$, hence Theorem 2.1 follows. ■

Proof of Theorem 2.2 Define $\Lambda_0 = (I_n - A_0)^2$ and $\Lambda_a = (I_n - A_a)^2$. Let RSS_a and RSS_0 be the residual sum of squares under the alternative and null hypotheses

$$\begin{aligned}
\text{RSS}_a &= (\hat{\mathbf{Z}}_a - \mathbf{Z})^{\mathrm{T}}(\hat{\mathbf{Z}}_a - \mathbf{Z}) = \mathbf{Z}^{\mathrm{T}}(I_n - A_a)^2\mathbf{Z} = \mathbf{Z}^{\mathrm{T}}\Lambda_a\mathbf{Z} \\
\text{RSS}_0 &= (\hat{\mathbf{Z}}_0 - \mathbf{Z})^{\mathrm{T}}(\hat{\mathbf{Z}}_0 - \mathbf{Z}) = \mathbf{Z}^{\mathrm{T}}(I_n - A_0)^2\mathbf{Z} = \mathbf{Z}_0^{\mathrm{T}}\Lambda_0\mathbf{Z} \\
\text{RSS}_0 - \text{RSS}_a &= \mathbf{Z}^{\mathrm{T}}(\Lambda_0 - \Lambda_a)\mathbf{Z} = \mathbf{Z}^{\mathrm{T}}\Lambda_\Delta\mathbf{Z}.
\end{aligned}$$

As shown in Lemma A.1 and Theorem 2.1, the bias is of a smaller order than the quadratic form $\boldsymbol{\varepsilon}^{\mathrm{T}}\Lambda_a\boldsymbol{\varepsilon}$. The residual sum square RSS_a can be approximated by $Q_a = \boldsymbol{\varepsilon}^{\mathrm{T}}\Lambda_a\boldsymbol{\varepsilon}$. A similar expression can be found for $\text{RSS}_0 - \text{RSS}_a$. Hence $Q_{\boldsymbol{\Delta}} = \boldsymbol{\varepsilon}^{\mathrm{T}}\Lambda_\Delta\boldsymbol{\varepsilon}$. Following Theorem 2.1, $Q_{\boldsymbol{\Delta}} \sim c_{\boldsymbol{\Delta}}\chi^2(b_{\boldsymbol{\Delta}})$ and $Q_a \sim c_a\chi^2(b_a)$ with parameters similar to Equation (A3).

Based on Craig's Theorem in Aitken (1950), the sufficient and necessary condition for the quadratic forms $\boldsymbol{\varepsilon}^{\mathrm{T}}C\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^{\mathrm{T}}D\boldsymbol{\varepsilon}$ to be statistically independent is that $CD = 0$. Although in our case the exact equality is not true, we will show that $\|\Lambda_a\Lambda_\Delta\|_\infty = o_p(1)$ where the max norm of a matrix $A_{n\times n}$ is defined as $\|A\|_\infty = \max_{1\le i,j\le n}\{|a_{ij}|\}$. Hence the two quadratic forms are asymptotically independent. Similar definitions about the asymptotic idempotent and asymptotic orthogonal can be found in Huang and Chen (2008).

With asymptotic independence, an asymptotic F-test can be constructed as

$$\frac{(\text{RSS}_0 - \text{RSS}_a)/(c_{\boldsymbol{\Delta}}b_{\boldsymbol{\Delta}})}{\text{RSS}_a/(c_{\mathbf{a}}b_{\mathbf{a}})} \sim F_{b_{\boldsymbol{\Delta}}, b_{\mathbf{a}}}.$$

Notice that $c_{\boldsymbol{\Delta}}b_{\boldsymbol{\Delta}} = \sigma_\varepsilon^2\text{trace}(\Lambda_\Delta)$, $c_{\mathbf{a}}b_{\mathbf{a}} = \sigma_\varepsilon^2\text{trace}(\Lambda_a)$, $b_\Delta = \text{trace}^2(\Lambda_\Delta)/\text{trace}(\Lambda_\Delta^2)$, and $b_a = \text{trace}^2(\Lambda_a)/\text{trace}(\Lambda_a^2)$. When defining $\nu_1, \nu_2, \delta_1, \delta_2$ as in Theorem 2.2, the asymptotic distribution in Equation (7) follows.

The rest is to show $\|\Lambda_a\Lambda_\Delta\|_\infty = o_p(1)$. The two smoothing matrices A_a and A_0 are both symmetric but not idempotent. However, they can be approximated by the projection matrix P_a in the sense that $\|A_a - P_a\|_\infty = o_p(1)$. Note that the matrix-induced norm is defined as $\|A_a - P_a\|_2 = \max[\{\|(A_a - P_a)\boldsymbol{\omega}\|^2/\|\boldsymbol{\omega}\|^2\}^{1/2}, \forall\boldsymbol{\omega} \neq 0 \in R^n]$. And for any matrix $A_{n\times n}$, we have $\|A\|_\infty \le \|A\|_2$.

Equation (A1) leads to

$$\begin{aligned}
\frac{\|(A_a - P_a)\boldsymbol{\omega}\|^2}{\|\boldsymbol{\omega}\|^2} &= \frac{\{\boldsymbol{\omega}^{\mathrm{T}}(A_a - P_a)^2\boldsymbol{\omega}\}}{\|\boldsymbol{\omega}\|^2} \\
&= \frac{\{\boldsymbol{\omega}^{\mathrm{T}}((1/n)G^*(1 + o_p(1)))^2\boldsymbol{\omega}\}}{\|\boldsymbol{\omega}\|^2} \\
&= \frac{\{n^{-2}\cdot\boldsymbol{\omega}^{\mathrm{T}}G^{*2}\boldsymbol{\omega}\}(1 + o_p(1))}{\|\boldsymbol{\omega}\|^2} \\
&= \frac{n^{-2}\text{trace}(\boldsymbol{\omega}^{\mathrm{T}}C_a A_n^{*-2}B_a^2(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}\boldsymbol{\omega})(1 + o_p(1))}{\|\boldsymbol{\omega}\|^2}.
\end{aligned}$$

Using arguments similar to those in Theorem 2.1, the last term is equal to $n^{-2}\text{trace}\{\boldsymbol{\omega}^{\mathrm{T}}C_a(C_a^{\mathrm{T}}C_a)^{-1}C_a^{\mathrm{T}}\boldsymbol{\omega}\}(1 + o_p(1))/\|\boldsymbol{\omega}\|^2$. It can be further simplified to $n^{-2}\|\boldsymbol{\omega}\|^2(1 + o_p(1))/\|\boldsymbol{\omega}\|^2 = n^{-2}\{1 + o_p(1)\}$ after a straightforward calculation. Hence, $\|A_a - P_a\|_\infty = O_p(n^{-1}) = o_p(1)$ as $\|(A_a - P_a)\boldsymbol{\omega}\|^2/\|\boldsymbol{\omega}\|^2 = O_p(n^{-2})$ holds for any nonzero vector $\boldsymbol{\omega}$. In addition, it can be proved that $\|(A_a - P_a)^l\|_\infty = O(n^{-l})$, for $l = 2, 3, 4$.

Recall that P_a and P_0 are two projection matrices. Hence $(I_n - P_a)C_a = 0$, $(I_n - P_0)C_0 = 0$. Because C_0 is a submatrix of C_a (as all its columns are from C_a), we have $(I_n - P_a)C_0 = 0$ and $(I_n - P_a)(I_n - P_0) = (I_n - P_a)$. Furthermore, $(I_n - P_a)^4 = (I_n - P_a)$ and $(I_n - P_a)^2(I_n - P_0)^2 = (I_n - P_a)$. It follows that

$$\begin{aligned}
\|\Lambda_a(\Lambda_0 - \Lambda_a)\|_\infty &= \|(I_n - A_a)^2(I_n - A_0)^2 - (I_n - A_a)^4\|_\infty \\
&\le \|(I_n - A_a)^2(I_n - A_0)^2 - (I_n - P_a)^2(I_n - P_0)^2\|_\infty \\
&\quad + \|(I_n - A_a)^4 - (I_n - P_a)^4\|_\infty.
\end{aligned}$$

The first term on the right-hand side is bounded by

$$\|(P_a - A_a)^2(I_n - P_0)^2\|_\infty + \|(P_a - A_a)^2(P_0 - A_0)^2\|_\infty$$
$$\leq n\|(P_a - A_a)^2\|_\infty\{\|(I_n - P_0)^2\|_\infty + \|(P_0 - A_0)^2\|_\infty\}.$$

The last step follows the fact that $\|A \cdot B\|_\infty \leq n \cdot \|A\|_\infty\|B\|_\infty$ for any two $n \times n$ matrices A and B. Because matrix P_0 is a projection matrix, then $\|\boldsymbol{\omega}\|^2 = \|(I_n - P_0)\boldsymbol{\omega}\|^2 + \|P_0\boldsymbol{\omega}\|^2$ and $\|(I_n - P_0)\boldsymbol{\omega}\|^2 \leq \|\boldsymbol{\omega}\|^2$, for any vector $\boldsymbol{\omega}$. It leads to $\|I_n - P_0\|_\infty \leq 1$. So $\|(I_n - A_a)^2(I_n - A_0)^2 - (I_n - P_a)^2(I_n - P_0)^2\|_\infty = O_p(n^{-1})$.

The second term is $\|(I_n - A_a)^4 - (I_n - P_a)^4\|_\infty = \|(P_a - A_a)^4\|_\infty = O_p(n^{-4})$. Hence, the sum of all terms leads to $\|\Lambda_a(\Lambda_0 - \Lambda_a)\|_\infty = O_p(n^{-1}) + O_p(n^{-4}) = o_p(1)$. Theorem 2.2 is proved. ■