Statistical inferences on partially linear models and their applications

Runze Li
Department of Statistics
Pennsylvania State University
University Park, PA 16802

Lei Nie
Department of Mathematics and Statistics
University of Maryland at Baltimore County
Baltimore, MD21250

December, 2004

Abstract

In this paper, we propose a class of partially nonlinear models, and develop two new estimation procedures for these models. Asymptotic normality of the resulting estimates are established. We further propose an estimation procedure and a generalized likelihood ratio test for the baseline function in partially nonlinear models. Asymptotic properties of the newly proposed estimation procedure and test are derived. Finite sample performance of the proposed inference procedures are assessed by Monte Carlo simulation studies. We further illustrate the proposed procedures by an application in the field of ecology.

Key Words: Generalized likelihood ratio test, partially linear models, profile likelihood.

Abbreviated Title: Partially Nonlinear Models

1 Introduction

Semiparametric regression models are useful for data analysis since they retain the flexibility of nonparametric models and the ease of interpretation of parametric models. Bickel, *et al.*(1993) gives insightful procedures to construct efficient estimators for parameters in the parametric component. Ruppert, Wand and Carroll (2003) presents various estimation procedures and many applications of semiparametric regression models. The current work proposes a new class of semiparametric models, called partially nonlinear models.

This work was motivated by a real data example in the field of ecology. Detailed analysis of this example will be given in Section 3.2. Here, we give only a brief introduction. It is known that sunlight intensity affects the rate of photosynthesis in an ecosystem. Since leaves absorb carbon dioxide (CO_2) during the course of photosynthesis, the *Net Ecosystem Exchange* of CO_2, denoted by NEE, is used to measure the level of photosynthetic activity in a natural ecosystem. Photosynthetic rate as measured by NEE is dependent on the amount of *Photosynthetically Active Radiation* available to an ecosystem, denoted by PAR.

It is believed based on empirical studies that the relationship between NEE and PAR is nonlinear and can be characterized by the following model

$$\mathrm{NEE} = R - \frac{\beta_1 \mathrm{PAR}}{\mathrm{PAR} + \beta_2} + \varepsilon, \tag{1.1}$$

where ε is random error with zero mean, and R, β_1 and β_2 are unknown parameters with physical interpretations. Specially, R is the dark respiration rate, β_1 is the light-saturated net photosynthetic rate, and β_1/β_2 is the apparent quantum yield. The empirical NEE-PAR relationship in (1.1) has been applied widely in assessing net primary productivity (NPP) since remote sensing data became available (see, for instance, Montheith, 1972, and more recent work by Ruimy et al., 1999 and references therein). The model in (1.1) was originally adopted based on laboratory experiments in which climate variables, such as temperature and moisture availability, can be well controlled. However, the temperature for a natural ecosystem cannot be controlled, and the parameter R most likely depends on the temperature. To demonstrate this, 2-dimensional kernel smoothing regression is used to estimate the regression function of the NEE on the PAR and the temperature using the data analyzed in Section 3.2. The lines in Figure 2(a) depict the estimated regression function of NEE on

PAR, given three different values of temperature. The nearly parallel pattern of the dashed lines suggests it may be appropriate to consider a partially nonlinear model

$$\mathrm{NEE} = R(\mathrm{T}) - \frac{\beta_1 \mathrm{PAR}}{\mathrm{PAR} + \beta_2} + \varepsilon,$$

where T is temperature.

Let Y is a response variable, and $\{\mathbf{x}, U\}$ be its covariates. In this paper, we consider a new class of semiparametric models:

$$Y = \alpha(U) + g(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon, \tag{1.2}$$

where $\alpha(\cdot)$ is an unknown smoothing function, $g(\cdot;\cdot)$ is a pre-specified function, $\boldsymbol{\beta}$ is an unknown parameter vector, and ε is a random error with mean zero and variance σ^2. Partially linear models, which are popular in the literature of semiparametric regression modeling (Engle, *et al.*, 1986, Heckman, 1986, Chen, 1988, Speckman, 1988 and among others), can be viewed as special cases of model (1.2). Härdle, Liang and Gao (2000) gave a systematic study for the partially linear models. In this paper, we concentrate on the nonlinear function $g(\cdot;\cdot)$ of $\boldsymbol{\beta}$. Thus, model (1.2) will be referred to as a partially nonlinear model. Model (1.2) retains the flexibility of nonparametric modeling for the baseline function and also retains the model interpretability of nonlinear parametric models. Nonlinear parametric regression models have been widely used in statistical literature. Many interesting examples and applications of such models are given in the reference books by Bates and Watts (1988), Seber and Wild (1989), and Huet (2004). Nonlinear modeling techniques have been also applied for econometric data and time series (Gallant,1987, and Fan and Yao, 2003).

We first propose an estimation procedure for $\boldsymbol{\beta}$ by using a profile nonlinear least squares technique. The consistency and asymptotic normality of the resulting estimate are established. Since the profile nonlinear least squares approach requires an iteration procedure to search for the solution, the procedure is computationally expensive. Therefore, we also propose an alternative estimation procedure using linear approximation. The linear approximation approach is computational less expensive, but shares the same asymptotic efficiency as the profile nonlinear least squares approach. We further propose an estimation procedure for the nonparametric baseline function $\alpha(\cdot)$. It may also be of interest to test whether the baseline function has some simple form (e.g, constant, or linear). To this end, we extend the generalized likelihood ratio type test (Fan, Zhang and Zhang, 2001) for the partially nonlinear model. We show that the proposed likelihood ratio test has a chi-square limiting distribution under the null hypothesis.

The paper is organized as follow. In Section 2, we first propose two estimation procedures for $\boldsymbol{\beta}$ by employing profile nonlinear least squares techniques and linear approximation. We further develop statistical inference procedures for the baseline function $\alpha(\cdot)$. In Section 3, we conduct numerical studies and assess finite sample performance of the newly proposed procedures. A real data example is illustrated the proposed procedures. Section 4 presents the regularity conditions and technical proofs.

2 Inference Procedures

Suppose that $\{\mathbf{x}_i, u_i, y_i\}, i = 1, \cdots, n$, is a random sample from the partially nonlinear model (1.2). Due to the curse of dimensionality, it is assumed throughout this paper that U is a univariate continuous random variable. In this section, we propose two estimation procedures for $\boldsymbol{\beta}$ and an inference procedure for the baseline function.

2.1 Profile nonlinear least squares approach

The profile nonlinear least squares method is in the same spirit as generalized profile likelihood principle (Severini and Wong, 1992). Profile likelihood approaches have been used to construct efficient estimates for partially linear models and generalized partially linear models (Severini and Staniswalis, 1994) Fan and Huang, 2004). In this paper, we propose profile nonlinear least squares method for to semiparametric nonlinear models (1.2).

Define a nonlinear least squares function

$$Q(\alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i - \alpha(u_i) - g(\mathbf{x}_i; \boldsymbol{\beta})\}^2. \tag{2.1}$$

For a given $\boldsymbol{\beta}$, let $y_i^* = y_i - g(\mathbf{x}_i; \boldsymbol{\beta})$. Then, $\{y_1^*, \cdots, y_n^*\}$ can be viewed as a sample from

$$y^* = \alpha(u) + \varepsilon. \tag{2.2}$$

This is a 1-dimensional ordinary nonparametric regression model. Thus, $\alpha(\cdot)$ can be easily estimated using any linear smoother. Here, we will employ local linear regression (Fan and

Gijbels, 1996). For any u in a neighborhood of u_0, it follows by Taylor's expansion that

$$\alpha(u) \approx \alpha(u_0) + \alpha'(u_0)(u - u_0) \widehat{=} a + b(u - u_0).$$

Let $K(\cdot)$ be a kernel function and h be a bandwidth. The local linear regression approach finds local parameters (a, b) to minimize

$$\sum_{i=1}^{n} \{y_i^* - a - b(u_i - u_0)\}^2 K_h(u_i - u_0),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. Note that local linear regression results in a linear estimator for $\alpha(\cdot)$ in terms of y_i^*. Let $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{\alpha} = (\alpha(u_1), \cdots, \alpha(u_n))^T$, and $\mathbf{g}(\boldsymbol{\beta}) = (g(\mathbf{x}_1; \boldsymbol{\beta}), \cdots, \mathbf{g}(\mathbf{x}_n; \boldsymbol{\beta}))^T$. Thus,

$$\hat{\boldsymbol{\alpha}} \widehat{=} \hat{\boldsymbol{\alpha}}(\cdot; \boldsymbol{\beta}) = S_h\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\}, \tag{2.3}$$

where S_h is an $n \times n$ smoothing matrix depending only on $u_1, \cdots, u_n$ and the bandwidth h.

Substituting $\hat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}$ in (2.1) results in *profile nonlinear least squares*:

$$Q(\boldsymbol{\beta}) \widehat{=} Q\{\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \boldsymbol{\beta}\} = \|(I_n - S_h)\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta})\}\|^2, \tag{2.4}$$

where I_n is the $n \times n$ identity matrix, and $\| \cdot \|$ stands for the Euclidean norm. Minimizing $Q(\boldsymbol{\beta})$ yields a nonlinear profile least squares estimator $\hat{\boldsymbol{\beta}}$. Furthermore, substituting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ in (2.3) yields

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}(\cdot; \hat{\boldsymbol{\beta}}) = S_h\{\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\beta}})\}. \tag{2.5}$$

Let $g'(\mathbf{x}, \boldsymbol{\beta})$ be $\partial g(\mathbf{x}; \boldsymbol{\beta})/\partial \boldsymbol{\beta}$, and $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$. Denote

$$\mathbf{A} = E[g'(\mathbf{x}; \boldsymbol{\beta}_0) - E\{g'(\mathbf{x}; \boldsymbol{\beta}_0)|U\}]^{\otimes 2}.$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$.

Theorem 1. *Suppose that the matrix* $\mathbf{A}$ *is finite positive definite. Under Conditions (A)—(F) given in Section 4,*

(a) with probability tending to one, there exists a minimizer $\hat{\boldsymbol{\beta}}_P$ *of* $Q(\boldsymbol{\beta})$ *in (2.4) such that* $\hat{\boldsymbol{\beta}}_P$ *is a consistent estimator for* $\boldsymbol{\beta}$*;*

(b)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_P - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 \mathbf{A}^{-1}),$$

where "$\xrightarrow{\mathcal{D}}$" stands for convergence in distribution.

Furthermore, $\text{cov}\{\hat{\boldsymbol{\beta}}_P\}$ can be consistently estimated by

$$\hat{\sigma}^2_{\varepsilon,P} \left[\sum_{i=1}^n \{g'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_P) - \bar{g}'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_P)\}^{\otimes 2} \right]^{-1}, \tag{2.6}$$

where $\bar{g}'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_P) = n^{-1} \sum_{i=1}^n g'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_P)$ and $\hat{\sigma}^2_{\varepsilon,P}$ is the residual mean squared error of the profile nonlinear estimator.

2.2 Linear approximation approach

Minimizing the profile nonlinear least squares (2.4) usually involves heavy computation. In this section, we propose a new estimation procedure which reduce the computational burden by linearly approximating the nonlinear function $g(\cdot;\cdot)$.

Let $\hat{\boldsymbol{\beta}}_I$ be a consistent estimate of $\boldsymbol{\beta}$. We will introduce an easy way to construct $\hat{\boldsymbol{\beta}}_I$ later on. Applying Taylor expansion for $g(\cdot;\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}_I$, we have

$$g(\mathbf{x};\boldsymbol{\beta}) = g(\mathbf{x};\hat{\boldsymbol{\beta}}_I) + g'(\mathbf{x};\hat{\boldsymbol{\beta}}_I)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_I) + o_P(\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_I\|).$$

Thus, it follows that

$$y_i \approx \alpha(u_i) + g(\mathbf{x}_i;\hat{\boldsymbol{\beta}}_I) + g'(\mathbf{x}_i;\hat{\boldsymbol{\beta}}_I)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_I) + \varepsilon_i \tag{2.7}$$

Let $z_i = y_i - g(\mathbf{x}_i;\hat{\boldsymbol{\beta}}_I) + g'(\mathbf{x};\hat{\boldsymbol{\beta}}_I)^T\hat{\boldsymbol{\beta}}_I$, and consider the following linear approximation model

$$z_i = \alpha(u_i) + g'(\mathbf{x}_i;\hat{\boldsymbol{\beta}}_I)^T\boldsymbol{\beta} + \varepsilon. \tag{2.8}$$

This is a partial linear model, so various existing estimation procedures can be directly employed to estimate $\boldsymbol{\beta}$. Here we use the profile least squares technique to derive an estimator for $\hat{\boldsymbol{\beta}}$. For a given $\boldsymbol{\beta}$, let $z_i^* = z_i - g'(\mathbf{x}_i;\hat{\boldsymbol{\beta}}_I)^T\boldsymbol{\beta}$, then

$$z_i^* = \alpha(u_i) + \varepsilon$$

which is 1-dimensional nonparametric regression model. Various nonparametric estimation procedures can be used to estimate $\alpha(\cdot)$. Here we will use local linear regression to estimate the baseline function. Denote $\mathbf{z}^* = (z_1^*, \cdots, z_n^*)^T$ and $\boldsymbol{\alpha} = (\alpha(u_1), \cdots, \alpha(u_n))^T$. Since the local linear regression is linear smoother, we have

$$\hat{\boldsymbol{\alpha}} = S_h \mathbf{z}^*$$

Substituting $\hat{\boldsymbol{\alpha}}$ for α in (2.8), then it follows that

$$(I - S_h)\mathbf{z} = (I - S_h)\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{z} = (z_1, \cdots, z_n)^T$ and $\mathbf{g}'(\hat{\boldsymbol{\beta}}_I) = (g'(\mathbf{x}_1; \hat{\boldsymbol{\beta}}_I), \cdots, g'(\mathbf{x}_n; \hat{\boldsymbol{\beta}}_I))^T$. Therefore, a least squares estimate for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_L = \{\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T (I - S_h)^T (I - S_h)\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^{-1}\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T (I - S_h)^T (I - S_h)\mathbf{z}.$$

Theorem 2. *Suppose that the matrix* $\mathbf{A}$ *is finite positive definite. Under Conditions (A)—(H) given in Section 4 and* $\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 \mathbf{A}^{-1})$$

as $n \to \infty$.

From Theorems 1 and 2, $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_L$ share the same asymptotic distribution. Similar to (2.6), $\text{cov}\{\hat{\boldsymbol{\beta}}_L\}$ can be consistently estimated by

$$\hat{\sigma}_{\varepsilon,L}^2 \left[\sum_{i=1}^n \{g'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_L) - \bar{g}'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_L)\}^{\otimes 2} \right]^{-1}, \tag{2.9}$$

where $\bar{g}'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_L) = n^{-1}\sum_{i=1}^n g'(\mathbf{x}_i; \hat{\boldsymbol{\beta}}_L)$ and $\hat{\sigma}_{\varepsilon,L}^2$ is the corresponding residual mean squared errors.

In practical implementation, one needs to obtain $\boldsymbol{\beta}_I$ first. We next propose an difference based estimator of $\boldsymbol{\beta}$ and demonstrate the resulting estimator has a root n consistency and asymptotic normality.

Reorder the data from the least to largest according to the value of the variable $\{u_i\}$. Denote $(\mathbf{x}_{(i)}, u_{(i)}, y_{(i)})$, $i = 1, \cdots, n$, to be the ordered sample. First of all, observe that for $i = 1, \cdots, n-1$

$$y_{(i+1)} - y_{(i)} = \alpha(u_{(i+1)}) - \alpha(u_{(i)}) + g(\mathbf{x}_{(i+1)}; \boldsymbol{\beta}) - g(\mathbf{x}_{(i)}; \boldsymbol{\beta}) + e_i, \tag{2.10}$$

where stochastic error $e_i = \varepsilon_{(i+1)} - \varepsilon_{(i)}$. Note that U is a univariate continuous random variable. Under some mild conditions, $u_{(i+1)} - u_{(i)} = O_P(1/n)$ so that the term $\alpha(u_{i+1}) - \alpha(u_i) = O_P(1/n)$. Therefore

$$y_{(i+1)} - y_{(i)} \approx g(\mathbf{x}_{(i+1)}; \boldsymbol{\beta}) - g(\mathbf{x}_{(i)}; \boldsymbol{\beta}) + e_i, \tag{2.11}$$

Thus, the nonlinear least squares approach can be utilized to estimate $\boldsymbol{\beta}$. In other words,

$$\hat{\boldsymbol{\beta}}_I = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n-1} \{y_{(i+1)} - y_{(i)} - g(\mathbf{x}_{(i+1)}; \boldsymbol{\beta}) + g(\mathbf{x}_{(i)}; \boldsymbol{\beta})\}^2.$$

In the presence of ties among observation times u_i, this linear approximation to $\alpha(u_{i+1}) - \alpha(u_i)$ still holds. For simplicity, we call the resulting estimate as the *difference based estimate*.

When U and ε are independent, $\varepsilon_{(i)}$ are independently and identically distributed to the distribution of ε_1. Therefore,

$$E[\{y_{(i+1)} - y_{(i)}\} - \{g(\mathbf{x}_{(i+1)}; \boldsymbol{\beta}) - g(\mathbf{x}_{(i)}; \boldsymbol{\beta})\}] = O_P(1/n)$$

This enables us to show that $\hat{\boldsymbol{\beta}}_I$ is root n consistent. See Li and Nie (2005) for detailed discussion on difference based estimate.

2.3 Inference procedures for the baseline function

Estimation of $\alpha(\cdot)$

For a given root n consistent estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we estimate $\alpha(\cdot)$ by smoothing the partial residuals $\{r_i = y_i - g(\mathbf{x}_i; \hat{\boldsymbol{\beta}}), i = 1, \cdots, n\}$ over u_i. Here $\hat{\boldsymbol{\beta}}$ can be either $\hat{\boldsymbol{\beta}}_P$, $\hat{\boldsymbol{\beta}}_L$ or $\hat{\boldsymbol{\beta}}_I$. We will utilize local linear regression (Fan and Gijbels, 1996) to estimate $\alpha(u)$. For any u in a neighborhood of u_0, it follows by Taylor's expansion that

$$\alpha(u) \approx \alpha(u_0) + \alpha'(u_0)(u - u_0) \mathrel{\widehat{=}} a + b(u - u_0).$$

Let $K(\cdot)$ be a kernel function and h be a bandwidth. Local linear regression method finds local parameters (a, b) to minimize

$$\sum_{i=1}^{n} \{r_i - a - b(u_i - u_0)\}^2 K_h(u_i - u_0), \tag{2.12}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. This results in a nonparametric fit $\hat{\alpha}(\cdot;\hat{\boldsymbol{\beta}})$. Denote by $(\hat{a}, \hat{b})$ the minimizer of (2.12). Then

$$\hat{\alpha}(u_0;\hat{\boldsymbol{\beta}}) = \hat{a}, \quad \text{and} \quad \hat{\alpha}'(u_0;\hat{\boldsymbol{\beta}}) = \hat{b}.$$

Note that $\hat{\boldsymbol{\beta}}$ is root n consistent. The rate of convergence for $\hat{\boldsymbol{\beta}}$ is faster than that of the nonparametric estimator. Since the errors in estimation of $\boldsymbol{\beta}$ are therefore negligible in the nonparametric estimation of α, the value of $\boldsymbol{\beta}$ can be regarded as known. Furthermore, we have the following theorem.

Theorem 3. *Suppose that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$. Under Conditions (A)—(C), if $h \to 0$ and $nh \to \infty$, then for $u_0 \in \Omega$, the support of U , we have*

$$\sqrt{nh}\{\hat{\alpha}(u_0;\hat{\boldsymbol{\beta}}) - \alpha(u_0) - \frac{1}{2}\alpha''(u_0)\int t^2K(t)dth^2\} \xrightarrow{\mathcal{D}} N(0, \sigma^2(u_0)),$$

where

$$\sigma^2(u_0) = \sigma_\varepsilon^2\{f(u_0)\}^{-1}\int K(u)^2\,du,$$

where $f(\cdot)$ is the marginal density of U.

The bias and variance expressions are the same as to those in Fan (1993). We omit the details of the proof.

Bandwidth selection

The theoretic optimal bandwidth will be the same as that for 1-dimensional nonparametric model (2.8) regarding $\boldsymbol{\beta}$ is known since the expressions of asymptotic bias and variance are the same as those in Fan (1993). Thus, the bandwidth can be easily selected by data-driven method. In practice, let $y_i^* = y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_I)$, and then apply various existing bandwidth selection procedures such as the plug-in bandwidth selector (Ruppert, Sheather and Wand, 1995) for $(u_i, y_i^*), i = 1, \cdots, n$ to select a bandwidth h. This approach will be implemented in Section 3.2.

Generalized likelihood ratio tests on the baseline function

In practice, a natural question that may arise is whether $\alpha(\cdot)$ has a pre-specified parametric form. This leads us to consider the following hypothesis testing problem:

$$H_0 : \alpha(\cdot) = \alpha_0(\cdot, \boldsymbol{\theta}) \quad \text{versus} \quad H_1 : \alpha(\cdot) \neq \alpha_0(\cdot, \boldsymbol{\theta}),$$

where $\alpha_0(\cdot, \boldsymbol{\theta})$ is a pre-specified parametric function and the vector $\boldsymbol{\theta}$ consists of unknown parameters. For example, if we are interested in testing whether $\alpha(\cdot)$ really depends on U variable, then we may consider the null hypothesis that $\alpha(\cdot) = \alpha_0$, where α_0 is an unknown constant. This kind of hypothesis testing is referred to as nonparametric goodness of fit test. Fan, Zhang and Zhang (2001) proposed a generalized likelihood ratio (GLR) test to deal with this issue for nonparametric regression models. In this section, we propose a GLR test for the partially nonlinear model. For simplicity of presentation, we consider only test of linearity for $\alpha(\cdot)$:

$$H_0 : \alpha(u) = \alpha_0 + \alpha_1 u, \quad \text{versus} \quad H_1 : \alpha(u) \neq \alpha_0 + \alpha_1 u, \tag{2.13}$$

where α_0 and α_1 are unknown constants. Note that the parameter space under the null hypothesis is finite dimensional, while it is infinite dimensional under the alternative hypothesis. Thus, many traditional tests, such as likelihood ratio test, cannot be directly applied for the above hypothesis. Here we propose the GLR test to deal with this issue. To gain more insights into the construction of generalized likelihood ratio tests, assume, tentatively, that the random error ε is $N(0, \sigma^2)$. Then the likelihood function of the data $\{u_i, \mathbf{x}_i, y_i\}, i = 1, \cdots, n$, is proportional to

$$(\sigma_\varepsilon^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}[y_i - \alpha(u_i) - g(\mathbf{x}_i; \boldsymbol{\beta})]^2\right\}.$$

Let $\hat{\boldsymbol{\beta}}$ be a root n consistent estimate of $\boldsymbol{\beta}$ under H_1, $(\hat{\alpha}_0, \hat{\alpha}_1)^T$ be an estimate of $(\alpha_0, \alpha_1)^T$ under H_0 and $\hat{\alpha}(\cdot)$ be an estimate of $\alpha(\cdot)$ under H_1. Denote $\mathrm{RSS}(H_0)$ and $\mathrm{RSS}(H_1)$ to be the residual sum of squares under H_0 and H_1, respectively. That is, $\mathrm{RSS}(H_0) = \sum_{i=1}^{n}\{y_i - \hat{\alpha}_0 - \hat{\alpha}_1 u_i - g(\mathbf{x}_i; \hat{\boldsymbol{\beta}})\}^2$, and $\mathrm{RSS}(H_1) = \sum_{i=1}^{n}\{y_i - \hat{\alpha}(u_i) - g(\mathbf{x}_i; \hat{\boldsymbol{\beta}})\}^2$. A GLRT statistic is defined as

$$(n/2)\log\{\mathrm{RSS}(H_0)/\mathrm{RSS}(H_1)\}. \tag{2.14}$$

Under H_0, $\mathrm{RSS}(H_0) \approx \mathrm{RSS}(H_1)$, and the GLR test statistic is asymptotically equivalent to

$$\mathrm{GLRT}_0 = (n/2)\{\mathrm{RSS}(H_0) - \mathrm{RSS}(H_1)\}/\mathrm{RSS}(H_1). \tag{2.15}$$

Intuitively, under H_0, there will be little difference between $\mathrm{RSS}(H_0)$ and $\mathrm{RSS}(H_1)$. However, under the alternative hypothesis, $\mathrm{RSS}(H_0)$ should become systematically larger than

RSS(H_1), and hence the test statistic GLRT_0 will tend to take a large positive value. Hence, a large value of the test statistic GLRT_0 indicates that the null hypothesis should be rejected. Both test statistics in (2.14) and (2.15) are monotone functions of $\mathrm{RSS}(H_0)/\mathrm{RSS}(H_1)$. Therefore, they have the same power. However, GLRT_0 does not rely on normality assumption on the random error ε. Furthermore, we can show that the asymptotic null distribution of GLRT_0 is a chi-square distribution in the following theorem.

Theorem 4. *Suppose that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2})$, and $E|\varepsilon|^4 < \infty$. Under H_0 in (2.13) and Conditions (A)—(C) in the Appendix, if $h \to 0$ in such a way that $nh^{3/2} \to \infty$, then the GLR test statistic, defined by $r_K GLRT_0$, has an asymptotic χ^2 distribution with δ_n degrees of freedom, where $\delta_n = r_K|\Omega|\{K(0) - 0.5\int K^2(u)\,du\}/h$, and $|\Omega|$ stands for the length of the support of U, and $r_K = \{K(0) - 0.5\int K^2(t)\,dt\}/\int\{K(t) - 0.5K * K(t)\}^2\,dt$. Here $K * K$ stands for the convolution of K and K. The value of r_K for some commonly used kernel functions are listed below.*

Kernel	*Uniform*	*Epanechnikov*	*Biweight*	*Triweight*	*Gaussian*
r_K	*1.2000*	*2.1153*	*2.3061*	*2.3797*	*2.5375*

The above theorem extends the generalized likelihood ratio theory (Fan, Zhang and Zhang, 2001) for semiparametric nonlinear modeling, and unveils a new Wilks phenomenon for semiparametric inference: the asymptotic null distribution is a chi-square distribution and does not depend on the unknown parameters (α_0, β_0). We will also provide empirical justification to the null distribution. Similar to Cai, Fan and Li (2000), the null distribution of GLR test can be estimated using Monte Carlo simulation or a bootstrap procedure. This usually provides a better estimate than the asymptotic null distribution, since the degrees of freedom tend to infinity and the results in the above theorem give only the main order of the degrees of freedom.

Remark: Note that for a χ^2 random variable with degrees of freedom r, $\mathrm{Var}\{\chi^2(r)\} = 2E\{\chi^2(r)\}$. Thus, $r_K = 2E(\mathrm{GLRT}_0)/\mathrm{Var}(\mathrm{GLRT}_0)$. Using this relationship, we can derive a better approximation to the normalizing constant r_K using the bootstrap samples of GLRT_0. Specifically, let $\mathrm{mean}(\mathrm{GLRT}_0^*)$ and $\mathrm{var}(\mathrm{GLRT}_0^*)$ be the sample mean and the sample variance of bootstrap samples $\{\mathrm{GLRT}_{0,i}^*,\ i = 1, \cdots, N\}$ of GLRT_0, respectively. Then r_K can be replaced by $c_K = 2\,\mathrm{mean}(\mathrm{GLRT}_0^*)/\mathrm{var}(\mathrm{GLRT}_0^*)$. This will be implemented in Section 3.

3 Numerical study and Application

In this section, we assess the finite sample performance by Monte Carlo simulation. We further illustrate the proposed methodology by an application to the real data example introduced in Section 1. All simulations were conducted using SAS codes. In our simulation, the kernel function is taken to be the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$.

3.1 Monte Carlo simulation

We generate random samples from the following model

$$Y = \alpha(U) + g(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon,$$

where $\varepsilon \sim N(0,1)$, and $U \sim U(0,1)$, the uniform distribution over [0,1]. In our simulation, we consider the following two baseline functions:

$$\alpha_1(U) = \sin(2\pi U), \quad \text{and} \quad \alpha_2(U) = \exp(U),$$

and two nonlinear functions $g(\cdot;\cdot)$:

$$g_1(x; \beta_1, \beta_2) = -\beta_1 x/(x + \beta_2)$$

with $\beta_1 = 18$ and $\beta_2 = 0.8$, which were chosen to be close the estimate for the real data example in Section 3.2, and $x \sim N(0,1)$; and

$$g_2(x_1, x_2; \beta_1, \beta_2) = 20 \exp(\beta_1 x_1 + \beta_2 x_2)/\{1 + \exp(\beta_1 x_1 + \beta_2 x_2)\}$$

with $\beta_1 = \beta_2 = 1$, The covariate vector $(x_1, x_2)^T$ was simulated from a normal distribution with mean zero and $\text{cov}(x_i, x_j) = 0.5^{|i-j|}$. The sample size n was set to 200 or 400. For each case, we conduct 1000 Monte Carlo simulation.

Performance of $\hat{\boldsymbol{\beta}}$

Simulation results for $\hat{\boldsymbol{\beta}}$ are summarized in Table 1, in which "mean" stands for the average of the 1000 estimates of β's, and "SD" is the standard deviation of the 1000 estimates of β's and can be regarded as the true standard error of β's. "SE" and "Std(SE)" are the average and the standard deviation of the 1000 estimated standard errors using (2.6) and

(2.9). From Table 1, we can see the performance of the two estimation procedures for $\boldsymbol{\beta}$ are almost the same, and both perform well. Comparing with the standard deviation of SE's, the difference between SD and its corresponding SE is small. This indicates that the proposed formulae (2.6) and (2.9) work very well. Results in Table 1 correspond to the bandwidth $h = 0.1$, which equals approximately the optimal bandwidth selected by the plug-in method of Ruppert, Sheather and Wand (1995). To study the effects of choice of bandwidth, we consider a smaller bandwidth $h = 0.05$. Simulation results for β_1 in $g = g_2$ are summarized in Table 2. Results for other cases are similar and omitted here. From Tables 1 and 2, we can see that the proposed procedures work well for a large range of bandwidth.

Performance of $\hat{\alpha}(\cdot)$

The performance of $\hat{\alpha}(\cdot)$ is assessed by the square root of average squared errors (RASE),

$$\mathrm{RASE}^2 = n_{\mathrm{grid}}^{-1} \sum_{k=1}^{n_{\mathrm{grid}}} \{\hat{\alpha}(u_k) - \alpha(u_k)\}^2,$$

where $\{u_k, k = 1, \cdots, n_{\mathrm{grid}}\}$ are the grid points at which $\hat{\alpha}(\cdot)$ is evaluated. In our simulation, the bandwidth was set to 0.1 or 0.2. The sample averages of the RASEs based on 1000 replicates are listed in Table 3, which reports only the case with $g(x; \beta) = -\beta_1 x/(x + \beta_2)$. Results for the other case is similar, and are omitted here to save space.

Performance of generalized likelihood ratio test

We first verify empirically that the GLR test statistic defined in Section 2 has a chi-square distribution under the null hypothesis. To this end, we take $H_0 : \alpha(u) = a_0 + a_1 u$. where $(a_0, a_1) = \mathrm{argmin}_{c_0,c_1} E\|\alpha(u) - c_0 - c_1 u\|^2$. Such choice of (a_0, a_1) will pose the challenge to the GLR test in the power calculation below. Then we find the null distribution based on 1000 bootstrap simulations. We present only results for $\alpha(u) = \alpha_1(u)$. Results for $\alpha_2(u)$ are similar. Figure 1 (a) and (c) depict the estimated density of the null distribution of the GLR test statistic $c_K\mathrm{GLRT}_0$ for $g = g_1$ and g_2, respectively. We also plot the density of the chi-square distribution in order to examine whether the null distribution is close to a chi-square distribution. The degree of freedom is chosen to be the closest integer to the sample average of $c_K\mathrm{GLRT}_0$ values from the 1000 bootstrap samples. It is clear from Figure 1 (a) and (c) that the Wilks type results hold, i.e., the chi-squared distribution is a good approximation for the asymptotic behavior of $c_K\mathrm{GLRT}_0$ under H_0.

To examine the power of the proposed nonparametric goodness of fit test, we evaluate the power of the GLR test for the alternative model

$$\alpha(u) = (1 - c)(a_0 + a_1 u) + c\alpha(u),$$

for each given c, where $\alpha(u)$ equals either $\alpha_1(\cdot)$ or $\alpha_2(\cdot)$. We took $c = 0.2,\ 0.4,\ 0.6$. Results for $\alpha_2(\cdot)$ are similar to those of $\alpha_1(\cdot)$. We present only results for $\alpha_1(\cdot)$.

Figure 1 (b) and (d) depicts the three power functions based on 400 Monte Carlo simulations for the sample size, $n = 200$, at three different significance levels: 0.10, 0.05 and 0.01. The powers at $c = 0$ for the foregoing three significance levels are: 0.095, 0.043 and 0.018 for Figure 1 (b) and 0.090, 0.048 and 0.018 for Figure 1 (d), respectively. This shows that the bootstrap method gives us an approximately right sized test.

3.2 An Application

In this section, we illustrate the proposed methodology by an analysis of a real data set in the field of ecology. Of interest in this example is to study how temperature affects the relationship between the net ecosystem-atmosphere exchange of CO_2 (NEE) and the photosynthetically active radiation (PAR). This data set consists of 1997 observations of NEE, PAR and temperature (T), and was collected over a subalpine forest at approximately 3050 meter elevation above sea level by using three-dimensional sonic anemometers on hundreds of meter towers during parts of the growth season of 1999.

We take the NEE as the response variable, and PAR and T as covariates. We first consider a fully nonparametric regression model:

$$\text{NEE} = m(\text{PAR}, \text{T}) + \varepsilon,$$

where $m(\cdot, \cdot)$ is an unspecified smoothing function, and ε is random error with mean 0 and variance σ^2. Two dimensional kernel regression was used to estimate the regression $m(\cdot, \cdot)$. To examine how temperature affects the parameters in Model (1.2), we plot $\hat{m}(\text{PAR}, \text{T})$ versus PAR for given values of T. The three lines in Figure 2 (a) depict the plot of $\hat{m}(\text{PAR}, \text{T})$ over PAR for $T = 10.76, 13.29$ and 15.41, which correspond to the three sample quartiles of temperatures. From Figure 2 (a), we can see the nonlinear relationship between PAR and

NEE when temperature is fixed. The nearly parallel pattern of the three lines suggests that the parameters β_1 and β_2 do not vary over temperature, while different intercepts for the three lines show the dark respiration rate R changes over temperature. The monotone decreasing pattern of the lines implies that Model (1.2) may be appropriate when temperature is fixed.

From the above analysis, we consider a partially nonlinear model

$$\text{NEE} = R(T) - \frac{\beta_1 \text{PAR}}{\text{PAR} + \beta_2} + \varepsilon. \tag{3.1}$$

The proposed estimation procedure for $\boldsymbol{\beta}$ is used to fit the data set using Model (3.1). We first compute the difference based estimate $\hat{\boldsymbol{\beta}}_I$, and then use the plug-in method of Ruppert, Sheater and Wand (1995) to select a bandwidth. The resulting optimal bandwidth is $h = 1.286$. The resulting estimates using this bandwidth are given in Table 4 along with their standard errors

As proposed in Section 2.2, we estimate the baseline function based on the partial residuals, which are depicted in Figure 2(b), from which we can see that the partial residuals have an increasing trend over temperature. Figure 2(b) also depicts the estimated baseline function using local linear regression.

From Figure 2(b), the overall increasing trend might suggest a simple linear model for $R(T)$: $R(T) = a + b \times T$. Thus, it is of interest to test $H_0 : R(T) = a + b \times T$. Using nonlinear least squares approach, we obtain the resulting estimate for parameters in model (3.1) under the H_0 given by $R(T) = -2.7933 + 0.5535 \times T$, $\tilde{\beta}_1 = 17.5564$ and $\tilde{\beta}_2 = 0.7979$. The estimate of β_1 and β_2 is very close to the ones obtained under H_1 and listed in Table 4. This implies that the model under H_0 fits the data almost as well as the one under H_1, and indicates that test of linearity for this data set is challenging. We next compute the generalized likelihood ratio test GLRT_0 which equals 34.46 with P-value < 0.001 obtained by using 1000 bootstraps. The null distribution is depicted in Figure 2(c). Thus, the dark respiration rate is not linear in temperature. This example also demonstrate that the GLRT has good power.

4 Regularity Conditions and Proofs

Regularity Conditions

(A) The random variable U has a bounded support $\mathcal{U}$. Its density function $f(u)$ is Lipschitz continuous and bounded from 0 on its support.

(B) The true unknown smoothing function $\alpha_0(u)$ has a continuous second derivative.

(C) $K(u)$ is a positive, bounded, and symmetric function with compact support. Furthermore, $K(u)$ satisfies the Lipschitz condition. The functions $u^3K(u)$ and $u^3K'(u)$ are bounded and $\int u^4K(u)\,du < \infty$.

(D) $nh^8 \to 0$ and $nh^2/\{\ln(h)\}^2 \to \infty$.

(E) For any $\mathbf{x}$, $g(\mathbf{x}, \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ and the second derivatives of $g(\mathbf{x}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are continuous, $\boldsymbol{\beta} \in \mathcal{B}$, where $\mathcal{B}$ is a compact set.

(F) Let d be the dimension of $\boldsymbol{\beta}$, and

$$g'(\mathbf{x}_i, \boldsymbol{\beta}) = [\partial g(\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}]_{d\times 1}\,, \text{ and } g''(\mathbf{x}_i, \boldsymbol{\beta}) = \left[\partial^2 g(\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\right]_{d\times d},$$

and $\mathrm{Vech}\{\mathrm{g}''(\mathbf{x}, \boldsymbol{\beta})\}$ is the $d \times (d+1)/2$-vector of all second derivatives of $g(\mathbf{x}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. $\mathrm{E}\{g'(\mathbf{x}, \boldsymbol{\beta})\}^{\otimes 2}$, $\mathrm{E}[\mathrm{E}\{g'(\mathbf{x}, \boldsymbol{\beta})|U\}^{\otimes 2}]$, and $\mathrm{E}(\mathrm{E}[\{\mathrm{Vech}\{\mathrm{g}''(\mathbf{x}, \boldsymbol{\beta})\}|\mathrm{U}]^{\otimes 2})$ are bounded in a neighborhood of $\boldsymbol{\beta}_0$.

(G) $\mathrm{E}\{\|g'(\mathbf{x}, \boldsymbol{\beta})\|^4\} < \infty$, $\mathrm{E}[\|\mathrm{Vech}\{\mathrm{g}''(\mathbf{x}, \boldsymbol{\beta})\}\|^4] < \infty$.

(H) $\|\mathrm{Vech}\{g''(x, \boldsymbol{\beta})\}\| \le B(\mathbf{x})$ for all $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$ and $\mathrm{E}\{\|B(\mathbf{x})\|^4\} < \infty$.

We now state Proposition 4 in Marc and Silverman (1982) as a lemma.

Lemma 1. *Let (X_i, Y_i), $i = 1, \cdots, n$, be i.i.d bivariate random variables, with density function $f(x, y)$. If $E|y|^s < \infty$, $\sup_x \int |y|^s f(x, y)\,dy < \infty$, and if $n^{2\delta-1}h \to \infty$ for some $\delta < 1 - \frac{1}{s}$, then*

$$\sup_x \left| \frac{1}{n}\sum_{i=1}^n [K_h(x_i - x)Y_i - E\{K_h(X_i - x)Y_i\}] \right| = O_p\left[\left\{\frac{-\ln(h)}{nh}\right\}^{1/2}\right]. \tag{4.1}$$

Lemma 2. *Let $\mathbf{g}'(\boldsymbol{\beta}) = (g'(\mathbf{x}_1, \boldsymbol{\beta}), \cdots, g'(\mathbf{x}_n, \boldsymbol{\beta}))^T$. Under conditions (A)-(G),*

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\boldsymbol{\beta}_0) = A\{1 + o_p(1)\}, \tag{4.2}$$

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T(I_n - S_h)^T(I_n - S_h)\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta}_0)\} = \xi_n + O_p(c_n^2), \tag{4.3}$$

where $c_n = \{\frac{-\ln(h)}{nh}\}^{1/2} + h^2$, $\xi_n = n^{-1}\sum_{i=1}^n [g'(\mathbf{x}_i;\boldsymbol{\beta}_0) - E\{g'(\mathbf{x};\boldsymbol{\beta}_0)|U = u_i\}]\varepsilon_i$.

Proof. Let

$$D_u = \begin{pmatrix} 1 & \frac{u_1-u}{n} \\ . & . \\ . & . \\ . & . \\ 1 & \frac{u_n-u}{n} \end{pmatrix}_{n\times 2}, \tag{4.4}$$

and $W_u = \mathrm{diag}\{K_h(u_1 - u), \cdots, K_h(u_n - u)\}$. By definition

$$S_h = \begin{pmatrix} (1,0)(D_{u_1}^T W_{u_1} D_{u_1})^{-1} D_{u_1}^T W_{u_1} \\ . \\ . \\ . \\ (1,0)(D_{u_n}^T W_{u_n} D_{u_n})^{-1} D_{u_n}^T W_{u_n} \end{pmatrix}_{n\times n}. \tag{4.5}$$

Using Lemma 1, and under conditions (A), (D), (E) and (G), we can show

$$(1,0)(D_u^T W_u D_u)^{-1} D_u^T W_u \mathbf{g}'(\boldsymbol{\beta}) = E\{g'(x,\boldsymbol{\beta})|U = u\}\{1 + O_p(c_n)\}, \tag{4.6}$$

holds uniformly in $u \in \mathcal{U}$. This implies, under condition (E) and (F) and by the weak law of large number,

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta})^T (I_n - S_h)^T (I_n - S_h)\mathbf{g}'(\boldsymbol{\beta}) = E[g'(x,\boldsymbol{\beta}) - E\{g'(x,\boldsymbol{\beta})|U\}]^{\otimes 2}\{1 + o_P(1) + O_P(c_n)\}, \tag{4.7}$$

which leads to (4.2) as $c_n \to 0$ under condition (D). We next show (4.3). Recall $\boldsymbol{\alpha}_0 = (\alpha_0(u_1), \cdots, \alpha_0(u_n))^T$, where $\alpha_0(u)$ is the true unknown smoothing function.

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T (I_n - S_h)^T (I_n - S_h)\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta}_0)\} = \frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T (I_n - S_h)^T (I_n - S_h)(\epsilon + \boldsymbol{\alpha}_0),$$

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$. Similar to (4.6),

$$(1,0)(D_u^T W_u D_u)^{-1} D_u^T W_u \boldsymbol{\alpha}_0 = \alpha_0(u)\{1 + O_p(c_n)\}$$

holds uniformly in $u \in \mathcal{U}$. This together with (4.6) implies that

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T (I_h - S_h)^T (I_h - S_h)\boldsymbol{\alpha}_0 = O_p(c_n^2 + c_n/\sqrt{n}) = O_P(c_n^2), \tag{4.8}$$

holds uniformly in $u \in \mathcal{U}$ as $\sqrt{n}c_n \to \infty$. Similarly,

$$(1,0)(D_u^T W_u D_u)^{-1} D_u^T W_u \epsilon = O_p(c_n)$$

holds uniformly in $u \in \mathcal{U}$ since $E(\varepsilon|U) = 0$. Thus, under condition (A), (B), (D), (E), and (G), it follows by the same argument that

$$\frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T (I_n - S_h)^T (I_n - S_h)\epsilon = \xi_n + O_P(c_n^2) \tag{4.9}$$

holds uniformly in $u \in \mathcal{U}$. Thus, (4.3) follows by (4.8) and (4.9). □

4.1 Proof of Theorem 1

We first show the consistency of $\hat{\boldsymbol{\beta}}_P$. It suffices to show that for any sufficiently small a

$$\lim_{n\to\infty} P\left[\sup_{\|\mathbf{t}\|=a} Q(\boldsymbol{\beta}_0 + \mathbf{t}) > Q(\boldsymbol{\beta}_0)\right] = 1. \tag{4.10}$$

Using the Taylor expansion,

$$Q(\boldsymbol{\beta}_0 + \mathbf{t}) - Q(\boldsymbol{\beta}_0) = Q'(\boldsymbol{\beta}_0)^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T Q''(\boldsymbol{\beta}^*)\mathbf{t}$$

where $\boldsymbol{\beta}^* = \lambda\boldsymbol{\beta}_0 + (1-\lambda)(\boldsymbol{\beta}_0 + \mathbf{t})$ for a certain $\lambda \in [0\ 1]$. Using (4.3)

$$n^{-1}Q'(\boldsymbol{\beta}_0) = -2n^{-1}\mathbf{g}'(\boldsymbol{\beta}_0)^T(I_n - S_h)^T(I_n - S_h)\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta}_0)\} = -2\xi_n + O_p(c_n^2). \tag{4.11}$$

Thus, $Q'(\boldsymbol{\beta}_0)^T\mathbf{t}$ is of the order $O_p\{\sqrt{n}(1 + \sqrt{n}c_n^2)\|\mathbf{t}\|\} = O_P(\sqrt{n}\|\mathbf{t}\|)$ as $\sqrt{n}c_n^2 \to 0$ by Condition (D).

$$\mathbf{t}^T Q''(\boldsymbol{\beta})\mathbf{t} = 2\mathbf{t}^T\mathbf{g}'(\boldsymbol{\beta})^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\boldsymbol{\beta})\mathbf{t} + G^T(\boldsymbol{\beta})(I_n - S_h)^T(I_n - S_h)(\epsilon + \boldsymbol{\alpha}_0 + \mathbf{d}),$$

where $\mathbf{d} = \mathbf{g}(\boldsymbol{\beta}) - \mathbf{g}(\boldsymbol{\beta}_0)$, $G(\boldsymbol{\beta}) = (G_1(\boldsymbol{\beta}), \cdots, G_n(\boldsymbol{\beta}))^T$, and $G_i(\boldsymbol{\beta}) = \mathbf{t}^T g''(\mathbf{x}_i, \boldsymbol{\beta})\mathbf{t}$. Thus, using (4.2), it can be shown that

$$\mathbf{t}^T Q''(\boldsymbol{\beta}^*)\mathbf{t} = 2n\{\mathbf{t}^T A\mathbf{t} + O_p(\|\mathbf{t}\|^3)\} \tag{4.12}$$

Since A is finite and positive definite, $\mathbf{t}^T Q''(\boldsymbol{\beta}^*)\mathbf{t}$ dominates $Q'(\boldsymbol{\beta}_0)^T\mathbf{t}$ for sufficiently large n and sufficiently small a. Hence (4.10) holds.

Now we show the asymptotic normality of $\hat{\boldsymbol{\beta}}$. Let $Q'_j(\boldsymbol{\beta})$ denote the j-th component of $Q'(\boldsymbol{\beta})$, and $Q''_j(\boldsymbol{\beta})$ be the j-row of $Q''(\boldsymbol{\beta})$. Using Taylor's expansion, for $j = 1, \cdots, d$,

$$0 = Q'_j(\hat{\boldsymbol{\beta}}) = Q'_j(\boldsymbol{\beta}_0) + Q''_j(\boldsymbol{\beta}^*_j)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{4.13}$$

where $\boldsymbol{\beta}^*_j$ lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Using (4.12), under conditions (A)-(H),

$$\frac{1}{2n} Q''(\boldsymbol{\beta}^*_j) = A_j\{1 + o_p(1)\},$$

in probability, where A_j is the j-row of A. Using (4.11), we have

$$\sqrt{n} A\{1 + o_p(1)\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sqrt{n}\xi_n\{1 + O_p(c_n^2)\} = \sqrt{n}\xi_n + o_P(1).$$

as $\sqrt{n}c_n^2 \to 0$ by Condition (D). Note that,

$$\xi_n = n^{-1}\sum_{i=1}^{n}[g'(\mathbf{x}_i; \boldsymbol{\beta}_0) - E\{g'(\mathbf{x}; \boldsymbol{\beta}_0)|U = u_i\}]\varepsilon_i.$$

Using the Slutsky theorem and the central limit theorem, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 A^{-1}).$$

4.2 Proof of Theorem 2

Let $\mathbf{g}(\boldsymbol{\beta}) = (g(\mathbf{x}_1, \boldsymbol{\beta}), \cdots, g(\mathbf{x}_n, \boldsymbol{\beta}))^T$. Note that

$$\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0 = \{\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^{-1}\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T(I_n - S_h)^T(I_n - S_h)\{\mathbf{z} - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\boldsymbol{\beta}_0\},$$

where $\mathbf{z} = (z_1, \cdots, z_n)^T$ (defined in Section 2.2) and $\mathbf{g}'(\hat{\boldsymbol{\beta}}_I) = (g'(\mathbf{x}_1; \hat{\boldsymbol{\beta}}_I), \cdots, g'(\mathbf{x}_n; \hat{\boldsymbol{\beta}}_I))^T$. We first show

$$\frac{1}{n}\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\hat{\boldsymbol{\beta}}_I) = A\{1 + o_p(1)\}. \tag{4.14}$$

Note that

$$\begin{aligned}
& \frac{1}{n}\mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\hat{\boldsymbol{\beta}}_I) \\
= \; & \frac{1}{n}\mathbf{g}'(\boldsymbol{\beta}_0)^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\boldsymbol{\beta}_0) \\
& + \frac{2}{n}\{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^T(I_n - S_h)^T(I_n - S_h)\mathbf{g}'(\boldsymbol{\beta}_0) \\
& + \frac{1}{n}\{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^T(I_n - S_h)^T(I_n - S_h)\{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}.
\end{aligned}$$

Under condition (A)-(H) and similar to the proof of Lemma 2, we can show that

$$\{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^T (I_n - S_h)^T (I_n - S_h) \mathbf{g}'(\boldsymbol{\beta}_0) = O_p(n\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\|)$$
$$\{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\}^T (I_n - S_h)^T (I_n - S_h) \{\mathbf{g}'(\boldsymbol{\beta}_0) - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\} = O_p(n\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\|^2).$$

Thus, (4.14) follows by (4.2) as $\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\| = O_p(1/\sqrt{n})$.

We next show that

$$\frac{1}{\sqrt{n}} \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)^T (I_n - S_h)^T (I_n - S_h)(\mathbf{z} - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\boldsymbol{\beta}_0) = \sqrt{n}\xi_n + o_P(1). \tag{4.15}$$

Using the definition of $\mathbf{z}$, we have

$$\mathbf{z} - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\boldsymbol{\beta}_0 = \mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\beta}}_I) + \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0)$$

and it follows by (4.3) that

$$\frac{1}{\sqrt{n}} \mathbf{g}'(\boldsymbol{\beta}_0)^T (I_n - S_h)^T (I_n - S_h)\{\mathbf{y} - \mathbf{g}(\boldsymbol{\beta}_0)\} = \sqrt{n}\xi_n + O_P(\sqrt{n}c_n^2) = \sqrt{n}\xi_n + o_P(1).$$

Thus, to establish (4.15), it is enough to show that

$$\frac{1}{\sqrt{n}} \mathbf{g}'(\boldsymbol{\beta}_0)^T (I_n - S_h)^T (I_n - S_h)\{\mathbf{g}(\boldsymbol{\beta}_0) - \mathbf{g}(\hat{\boldsymbol{\beta}}_I) + \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0)\} = o_p(1). \tag{4.16}$$

and

$$\frac{1}{\sqrt{n}} \{\mathbf{g}'(\hat{\boldsymbol{\beta}}_I) - \mathbf{g}'(\boldsymbol{\beta}_0)\}^T (I_n - S_h)^T (I_n - S_h)\{\mathbf{z} - \mathbf{g}'(\hat{\boldsymbol{\beta}}_I)\boldsymbol{\beta}_0\} = o_p(1), \tag{4.17}$$

By straightforward calculation, the left-hand side of (4.16) is of the order

$$O_P(\sqrt{n}\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\|^2) = O_P(1/\sqrt{n})$$

as $\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$. Furthermore, the left-hand side of (4.17) is of the order $O_P(c_n\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_0\|) = O_P(c_n/\sqrt{n})$. Thus, (4.15) holds.

Using (4.14), (4.15), it follows

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) = A\{1 + o_P(1)\}^{-1}\{\sqrt{n}\xi_n + o_P(1)\}.$$

By the Slutsky Theorem and the central limit theorem, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0, \sigma^2 A^{-1}).$$

This completes the proof of Theorem 2.

4.3 Proof of Theorem 4

Let $y_i^* = y_i - g(\mathbf{x}_i; \boldsymbol{\beta}_0)$. Thus,

$$y_i^* = \alpha(u_i) + \varepsilon_i.$$

Let $\tilde{\alpha}^*$ and $\hat{\alpha}^*(\cdot)$ be the estimate of α under H_0 and H_1, respectively. Denote $\mathrm{RSS}^*(H_0) = \sum_{i=1}^n (y_i^* - \tilde{\alpha}^*)^2$ and $\mathrm{RSS}^*(H_1) = \sum_{i=1}^n \{y_i^* - \hat{\alpha}^*(u_i)\}^2$. Define

$$\mathrm{GLRT}_0^* = (n/2)(\mathrm{RSS}^*(H_0) - \mathrm{RSS}^*(H_1))/\mathrm{RSS}^*(H_1)$$

By Theorem 5 of Fan, Zhang and Zhang (2001), it follows that

$$r_K \mathrm{GLRT}_0^* \overset{a}{\sim} \chi^2_{\delta_n}.$$

Because $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$, we can show that under H_0,

$$n^{-1}\{\mathrm{RSS}(H_0) - \mathrm{RSS}^*(H_0)\} = o_P(1), \text{ and } n^{-1}\{\mathrm{RSS}(H_1) - \mathrm{RSS}^*(H_1)\} = o_P(1).$$

This completes the proof.

Acknowledgements: The authors thank Dr. C. Yi for providing data analyzed in Section 3.2.

References

Bates, D. M., Watts, D. G. (1988). *Nonlinear regression Analysis and Its Applications*, Wiley, New York.

Bickel, P.J., Klaassen, A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semiparametric Models.* John Hopkins University Press, Baltimore.

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

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, **16**, 136-146.

Engle, R.F., Granger,C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Asso.*, **81**, 310-320.

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

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

Fan, J. and Huang, T. (2004). Profile likelihood inference on semiparametric varying-coefficient partially linear models. Manuscript.

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

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York.

Gallant, A. N. (1987). *Nonlinear Statistical Models*, Wiley, New York.

Härdle, W. Liang, H. and Gao, J. (2000). *Partially Linear Models*, Springer-Verlag, New York.

Heckman, N. (1986). Spline smoothing in partly linear models, *J. Royal Statist. Soc. B.* **48**, 244-248.

Huet, S. (2004). *Statistical Tools for Nonlinear Regression: a Practical Guide with S-PLUS and R Examples*, Springer, New York.

Li, R. and Nie, L. (2005). A new estimation procedure for partially nonlinear model via mixed effects approach. Manuscript.

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

Monteith, J.L. (1972). Solar radiation and productivity in tropical ecosystems, *Journal of Applied Ecology*, **9**, 747-766.

Ruimy, A., L. Kergoat, A. Bondeau, and The Participants of The Potsdam NpP. (1999). Model Intercomparison, Comparing global models of terrestrial net primary productivity (NPP): analysis of differences in light absorption and light-use efficiency, *Global Change Biology*, **5**, 56-64.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Jour. Amer. Statist. Assoc.*, **90**, 1257-1270.

Ruppert, D. Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, New York.

Seber, G.A.F. and Wild, C. J. (1989). *Nonlinear Regression*, Wiley, New York.

Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.*, **89**, 501–511.

Severini, T.A. and Wong, W.H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768-1802.

Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Royal Statist. Soc. B*, **50**, 413-436.

Table 1: Finite Sample Performance of $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_L$ with $h = 0.1$

		Profile Nonlinear LS			Linear Approximation		
β	(n, α, g)	Mean	SD	SE(Std(SE))	Mean	SD	SE(Std(SE))
$\beta_1 = 18$	(200, α_1, g_1)	18.294	1.524	1.417(0.452)	18.29	1.521	1.391(0.440)
	(400, α_1,g_1)	18.145	1.004	0.963(0.193)	18.144	1.004	0.954(0.191)
	(200, α_2,g_1)	18.294	1.524	1.416(0.452)	18.290	1.521	1.390(0.44)
	(400, α_2,g_1)	18.145	1.004	0.963(0.193)	18.144	1.004	0.953(0.191)
$\beta_2 = 0.8$	(200, α_1, g_1)	0.829	0.165	0.153(0.048)	0.829	0.164	0.150 (0.047)
	(400, α_1,g_1)	0.814	0.106	0.104(0.020)	0.814	0.106	0.103(0.020)
	(200, α_2,g_1)	0.829	0.165	0.153(0.048)	0.829	0.164	0.15(0.047)
	(400, α_2,g_1)	0.814	0.106	0.104(0.020)	0.814	0.106	0.103(0.020)
$\beta_1 = 1$	(200, α_1, g_2)	1.001	0.030	0.029(0.002)	1.001	0.030	0.029 (0.002)
	(400, α_1,g_2)	1.000	0.021	0.021(0.001)	1.000	0.021	0.021 (0.001)
	(200, α_2,g_2)	1.001	0.030	0.029(0.002)	1.001	0.029	0.029 (0.002)
	(400, α_2,g_2)	1.000	0.020	0.021(0.001)	1.000	0.020	0.021 (0.001)
$\beta_2 = 1$	(200, α_1, g_2)	1.000	0.029	0.029(0.002)	1.000	0.0029	0.0029(0.002)
	(400, α_1,g_2)	1.001	0.020	0.021(0.001)	1.001	0.020	0.021 (0.001)
	(200, α_2,g_2)	1.000	0.029	0.029(0.002)	1.000	0.0029	0.0029(0.002)
	(400, α_2,g_2)	1.001	0.020	0.021(0.001)	1.001	0.020	0.021 (0.001)

Table 2: Finite Sample Performance of $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_L$ with $h = 0.05$

		Profile Nonlinear LS			Linear Approximation		
	(n, α)	Mean	SD	SE(Std(SE))	Mean	SD	SE(Std(SE))
$\beta_1 = 1$	(200, α_1)	1.001	0.031	0.029(0.002)	1.000	0.031	0.029 (0.002)
	(400, α_1)	1.000	0.021	0.021(0.001)	1.000	0.021	0.021 (0.001)
	(200, α_2)	1.001	0.031	0.029 (0.002)	1.000	0.031	0.029 (0.002)
	(400, α_2)	1.001	0.021	0.021(0.001)	1.000	0.021	0.021 (0.001)

Table 3: Mean of RASEs for Baseline Function

	Profile Nonlinear LS		Linear Approximation	
(n, g)	$h = 0.1$	$h = 0.2$	$h = 0.1$	$h = 0.2$
(200, g_1)	0.123	0.117	0.123	0.116
(400, g_1)	0.056	0.058	0.056	0.058
(200, g_2)	0.039	0.032	0.039	0.032
(400, g_2)	0.0120	0.021	0.0120	0.021

Table 4: Estimated Coefficients and Their Standard Errors

Profile Nonlinear LS		Linear Approximation	
$\hat{\beta}_1(\mathrm{SE}(\hat{\beta}_1))$	$\hat{\beta}_2(\mathrm{SE}(\hat{\beta}_2))$	$\hat{\beta}_1(\mathrm{SE}(\hat{\beta}_1))$	$\hat{\beta}_2(\mathrm{SE}(\hat{\beta}_2))$
17.656 (0.400)	0.81 (0.072)	17.64 (0.384)	0.81 (0.063)

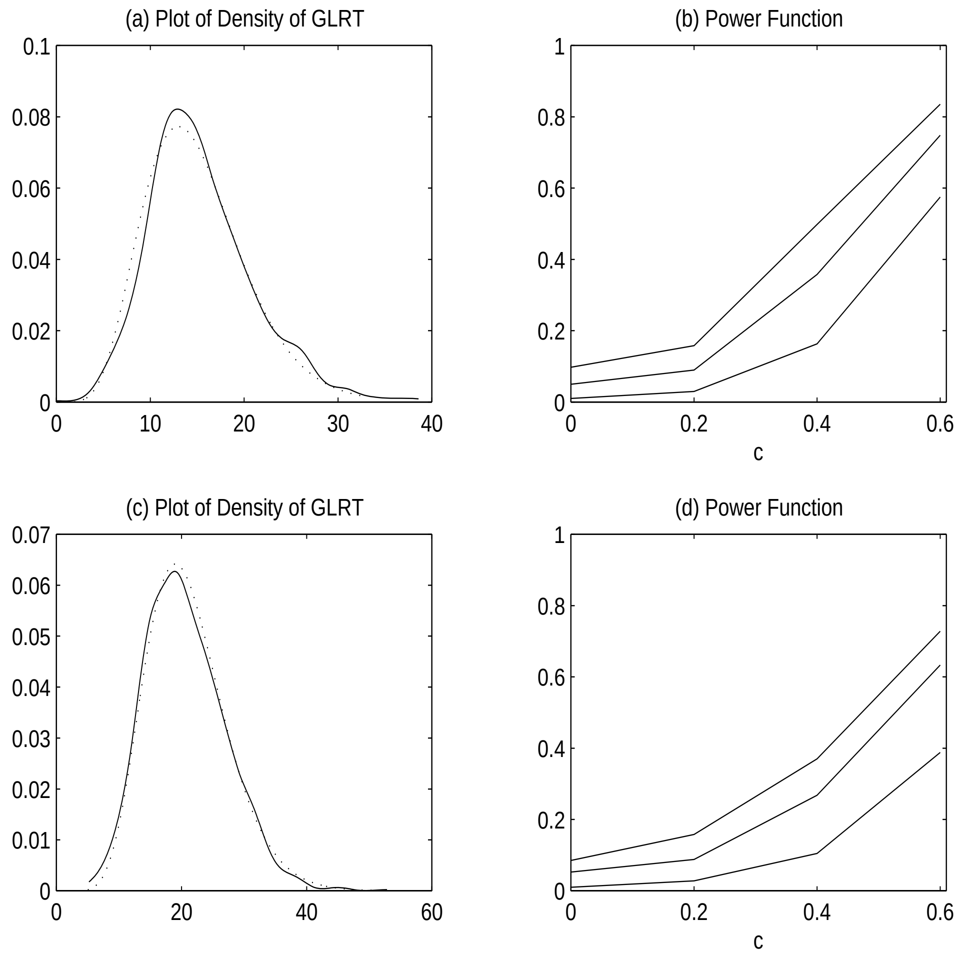

Figure 1: *Plot of Null Distribution of the Generalized Likelihood Ratio Test and its Power Function. (a) and (c) are the density of the generalized likelihood ratio test under the null hypothesis. The solid line is the estimated density curve using kernel density estimation and dash-dotted line is the density curve of a* χ^2*-distribution with degrees of freedom 15 and 21 for (a) and (c) respectively. (b) and (d) are power functions at level* α*=0.01, 0.05 and 0.10 from bottom to top, respectively.*

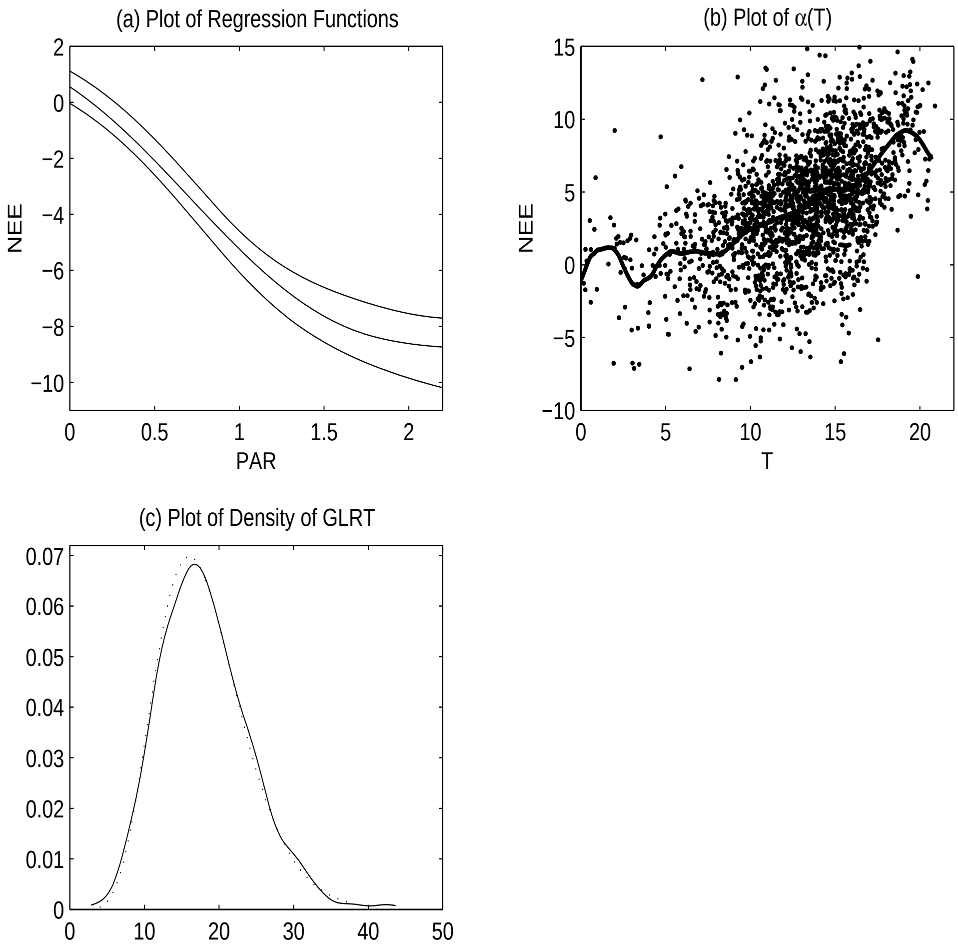

Figure 2: *Plots for Example in Section 3.2. The dashed lines in (a) are the fitted value of NEE versus PAR given temperature using 2-d kernel regression. From the bottom to top, the temperatures are 10.76, 13.29 and 15.41, which correspond to the three sample quartiles of temperatures. In (b), the solid line is an estimate of the baseline function, and the dots are partial residuals:* $y_i - \hat{\beta}_1 x_i/(x_i + \hat{\beta}_2)$. *In (c), the solid line is an estimate to the density of the GLRT under* H_0*, and dotted line is the density of a chi-square distribution with d.f. 18.*