Statistica Sinica **23** (2013), 1235-1255
doi:http://dx.doi.org/10.5705/ss.2012.075

SEMIPARAMETRIC ESTIMATION OF CONDITIONAL HETEROSCEDASTICITY VIA SINGLE-INDEX MODELING

Liping Zhu, Yuexiao Dong and Runze Li

Shanghai University of Finance and Economics, Temple University and Pennsylvania State University

Abstract: We consider a single-index structure to study heteroscedasticity in regression with high-dimensional predictors. A general class of estimating equations is introduced. The resulting estimators remain consistent even when the structure of the variance function is misspecified. The proposed estimators estimate the conditional variance function asymptotically as well as if the conditional mean function was given a priori. Numerical studies confirm our theoretical observations and demonstrate that our proposed estimators have less bias and smaller standard deviation than the existing estimators.

Key words and phrases: Conditional variance, heteroscedasticity, single-index model, volatility.

1. Introduction

Many scientific studies rely on understanding the local variability of the data, frequently characterized through the conditional variance in statistical modeling. The conditional variance plays an important role in a variety of statistical applications, such as measuring the volatility of risk in finance (Anderson and Lund (1997); Xia, Tong, and Li (2002)), monitoring the reliability of nonlinear prediction (Müller and Stadtmüller (1987); Yao and Tong (1994)), identifying homoscedastic transformations in regression (Box and Cox (1964); Carroll and Ruppert (1988)) and so on. Estimation of the conditional variance function is an important problem in statistics.

Let $Y \in \mathbb{R}$ be the response variable and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be the associated predictor vector. In this paper we study the conditional variance function of Y given $\mathbf{x}$, denoted $\mathrm{var}(Y \mid \mathbf{x})$. We write $E(Y \mid \mathbf{x})$ as the conditional mean function and $\varepsilon = Y - E(Y \mid \mathbf{x})$ as the random error. Then we have $\mathrm{var}(Y \mid \mathbf{x}) = E(\varepsilon^2 \mid \mathbf{x})$. Since ε is not observable, it is natural to replace the error with the residual $\widehat{\varepsilon} = Y - \widehat{E}(Y \mid \mathbf{x})$, where $\widehat{E}(Y \mid \mathbf{x})$ is an arbitrary consistent estimate of the conditional mean function. It is of interest to quantify the effect of this replacement on estimating the conditional variance function. This problem first received much attention when the predictor is univariate, $p = 1$. See, for

example, Hall and Carroll (1989), Wang et al. (2008), and references therein. Cai, Levine, and Wang (2009) generalized the results of Wang et al. (2008) to the case of multivariate $\mathbf{x}$. In their generalization, however, the nonparametric kernel regression was applied directly to estimate the multivariate regression function, which may not be effective due to the "curse of dimensionality". In regressions with univariate predictor and response, Ruppert et al. (1997) and Fan and Yao (1998) applied local linear regression to the squared residuals and demonstrated that such nonparametric estimate performs asymptotically as well as if the conditional mean were given a priori. Song and Yang (2009) derived asymptotically exact and conservative confidence bands for the heteroscedastic variance functions. Yin et al. (2010) extended the results of Fan and Yao (1998) to the case of multivariate response.

To model the conditional variance function when p is fairly large, we assume throughout that there exists a smooth function $\sigma^2(\cdot)$ and a $p \times 1$ vector $\boldsymbol{\beta}_0$ such that

$$\operatorname{var}(Y \mid \mathbf{x}) = \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}). \tag{1.1}$$

With an unspecified link function $\sigma^2(\cdot)$ and a univariate index $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$, (1.1) is both flexible and interpretable under the single-index structure. It provides a compromise between parametric models which are easily interpretable yet often too restrictive, and fully nonparametric models that are flexible but suffer from the "curse of dimensionality". For ease of presentation we assume that, for some unknown function $\ell(\cdot)$ and a $p \times 1$ vector $\boldsymbol{\alpha}_0$, the conditional mean function also admits the single-index structure

$$E(Y \mid \mathbf{x}) = \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}). \tag{1.2}$$

The assumption of model (1.2) is not essential and more general forms of the mean function may be assumed.

In this paper we propose a general class of estimating equations to estimate $\boldsymbol{\beta}_0$. By correctly specifying $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, the estimator of $\boldsymbol{\beta}_0$ is consistent even when the structure of the variance function $\sigma^2(\cdot)$ is misspecified. On the other hand, if the variance function $\sigma^2(\cdot)$ is correctly specified and estimated consistently, our proposed estimator of $\boldsymbol{\beta}_0$ is consistent without correctly specifying $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$.

The estimate of $\boldsymbol{\beta}_0$ from the estimating equations possesses an adaptive property: the proposed procedure estimates the conditional variance function as efficiently, asymptotically, as if the conditional mean were given a priori. This is achieved by replacing the error $\varepsilon = Y - E(Y \mid \mathbf{x})$ in the estimating equation with its corresponding residual $\widehat{\varepsilon} = Y - \widehat{E}(Y \mid \mathbf{x})$.

This paper is organized in the following way. In Section 2 we present the methodology and study its properties under both the population and the sample levels. In Section 3 we examine the finite sample performance of the proposed procedure through simulations and an application to a data set. This paper concludes with a brief discussion in Section 4. All technicalities are given in the Appendix.

2. A New Procedure and Its Theoretical Properties

2.1. A general class of estimating equations

Consider the estimating equation

$$E[\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\}\{\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\}'\mathbf{x}] = \mathbf{0}, \tag{2.1}$$

where $\varepsilon = Y - E(Y \mid \mathbf{x})$, and $\{\sigma^2(\cdot)\}'$ stands for the first order derivative of $\sigma^2(\cdot)$. Here (2.1) corresponds to the classical nonlinear least squares estimation of Ichimura (1993) and Härdle, Hall, and Ichimura (1993). A limitation of (2.1) is that it requires correct specification of $\sigma^2(\cdot)$ to obtain a consistent estimate for $\boldsymbol{\beta}_0$, and this may be troublesome in practice. To address this issue we propose a new class of estimating equations,

$$E\left[\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - \widetilde{E}\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\}\right] = 0, \tag{2.2}$$

where $\widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ and $\widetilde{E}\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)$ may be different from $\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ and $E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)$. When $\widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ and $\widetilde{E}\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right) = E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)$, (2.2) has the form

$$E\left[\left\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\}\right] = \mathbf{0}. \tag{2.3}$$

It is not difficult to verify that (2.3) produces a consistent estimate of $\boldsymbol{\beta}_0$. In other words, $\boldsymbol{\beta}_0$ is a solution to $E\left[\left\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right)\right\}\right] = \mathbf{0}$. An important virtue of (2.2) is that, as long as one of the functions $\widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ and $\widetilde{E}\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)$ is correctly specified, (2.2) yields a consistent estimate of $\boldsymbol{\beta}_0$. In particular, without knowing the exact form of $\sigma^2(\cdot)$, suppose we specify it as $\widetilde{\sigma}^2(\cdot)$, which may be different from $\sigma^2(\cdot)$, and impose the working estimating equation

$$E\left[\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\}\right] = \mathbf{0}. \tag{2.4}$$

Invoking a conditional expectation, (2.4) is equivalent to

$$E\left(E\left[\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\} \mid \mathbf{x}\right]\right) = \mathbf{0}.$$

To verify that this yields a consistent estimate of the true $\boldsymbol{\beta}_0$ at the population level, we recall that $E\left(\varepsilon^2 \mid \mathbf{x}\right) = \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ under (1.1). Then the left hand side of (2.4) can be further reduced as

$$\begin{aligned} & E\left[E\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) \mid \mathbf{x}\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\}\right] \\ & \quad = E\left[E\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right\}\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\}\right] \\ & \quad = E\left[\left\{\varepsilon^2 - \widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\} E\left\{\mathbf{x} - E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right\}\right], \end{aligned}$$

where the last term is obviously $\mathbf{0}$, indicating that (2.4) is able to produce a consistent estimator of $\boldsymbol{\beta}_0$. This derivation provides the motivation for (2.2), as it continues to yield a consistent estimate of $\boldsymbol{\beta}_0$ even if $\sigma^2(\cdot)$ is misspecified. As a special case, let $\widetilde{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = 0$ in (2.4) and get

$$E\left[\varepsilon^2\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\right] = \mathbf{0}, \tag{2.5}$$

which is similar to the estimating equation proposed by Li and Dong (2009) to recover the central solution space. In their context they utilize Y instead of ε^2 to estimate the mean function. We generalize the idea of the central solution space method to estimate the variance function. When $\mathbf{x}$ satisfies the linearity condition (Li (1991)), ($E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is a linear function of $\mathbf{x}$), then $\boldsymbol{\beta}_0$ must be proportional to $\{\mathrm{var}(\mathbf{x})\}^{-1}\mathrm{cov}(\mathbf{x}, \varepsilon^2)$. This property dramatically simplifies solving (2.5) at the sample level. Yet, unlike the response variable Y, the error term ε is not observable. This motivates us to examine the effect of estimating the mean function to obtain the residuals on estimating the variance function. We study this issue in the next section. Aside from this, we have seen that (2.5) is a specific member of the general class of estimating equations at (2.2).

On the other hand, correct specification of $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ may be challenging in some situations. Even if all components in $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ can be correctly specified, calculating $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is rather intensive when $\mathbf{x}$ is high dimensional. In order to simplify the calculation, suppose we estimate $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ by $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$. Then (2.2) becomes

$$E\left[\left\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - \widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\right] = \mathbf{0}, \tag{2.6}$$

noting that $E\left(\varepsilon^2 \mid \mathbf{x}\right) = \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ and conditioning on $\mathbf{x}$. Thus (2.2) continues to yield a consistent estimate of $\boldsymbol{\beta}_0$ even if $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is misspecified. For example, we can set $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = \mathbf{0}$ and then (2.6) becomes

$$E\left[\left\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\mathbf{x}\right] = 0. \tag{2.7}$$

We remark here that (2.5) and (2.7) are equivalent at the population level by noting that $E\left\{\varepsilon^2 E\left(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\right\} = E\left\{E\left(\varepsilon^2 \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right)\mathbf{x}\right\} = E\left\{\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\mathbf{x}\right\}$ under (1.1).

We have shown that the estimating equation at (2.2) has a desirable robustness property: as long as either $\widetilde{\sigma}^2(\cdot)$ or $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is correctly specified, the estimator based on the sample version of (2.2) is consistent. Both parametric and nonparametric methods could be used to model $\widetilde{\sigma}^2(\cdot)$ or $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$. For example, $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is typically assumed to be a linear function of $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ in the dimension reduction literature (Li (1991)). We propose to estimate $\widetilde{\sigma}^2(\cdot)$ and $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ via kernel regression. The theoretical properties of the sample estimates based on kernel regression are investigated in the next section.

2.2. Asymptotic properties

Given independent and identically distributed $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, we discuss sample versions of the proposed estimating equations and their asymptotic properties.

Ideally, one may estimate $\boldsymbol{\beta}_0$ by solving for $\widehat{\boldsymbol{\beta}}$ that satisfies

$$n^{-1/2}\sum_{i=1}^{n}\{\varepsilon_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \tag{2.8}$$

which is the sample counterpart of (2.3). In (2.8), the quantities $\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ and $\widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ are the consistent estimators of $\sigma^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ and $E(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$, respectively, and can be obtained through the classical kernel regression method. In practice, we have to replace the unobservable error ε_i with the residual $\widehat{\varepsilon}_i$. To guarantee the consistency of $\widehat{\boldsymbol{\beta}}$, we need consistent estimation of ε_i, indicating that we have to estimate $E(Y \mid \mathbf{x})$ consistently. Under (1.2), $E(Y \mid \mathbf{x}) = \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$, we estimate $\boldsymbol{\alpha}_0$ by solving for $\widehat{\boldsymbol{\alpha}}$ that satisfies

$$n^{-1/2}\sum_{i=1}^{n}\{Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \tag{2.9}$$

which is parallel to (2.8) with

$$\widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i) = \frac{\sum_{j=1}^{n} K_h(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\mathbf{x}_j}{\sum_{j=1}^{n} K_h(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)}, \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i) = \frac{\sum_{j=1}^{n} K_{h_1}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)Y_j}{\sum_{j=1}^{n} K_{h_1}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_j - \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)},$$

In this $K_h(\cdot) = K(./h)/h$ is the kernel function. h and h_1 are the bandwidths.

Next we calculate the residual $\widehat{\varepsilon}_i = Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)$, and get our final estimate of $\boldsymbol{\beta}_0$ by solving

$$n^{-1/2}\sum_{i=1}^{n}\{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\} = 0. \tag{2.10}$$

In (2.10) we estimate $E(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ and $\sigma^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ as

$$\widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)=\frac{\sum_{j=1}^n K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\mathbf{x}_j}{\sum_{j=1}^n K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)}, \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)=\frac{\sum_{j=1}^n K_{h_2}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\widehat{\varepsilon}_j^2}{\sum_{j=1}^n K_{h_2}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_j-\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)}.$$

To summarize, we implement the following algorithm to estimate $\boldsymbol{\beta}_0$.

1. Solve (2.9) to obtain $\widehat{\boldsymbol{\alpha}}$ through the following Newton-Raphson algorithm.
   (a) Start with an initial value $\boldsymbol{\alpha}^{(0)}$. In our implementation, we choose $\boldsymbol{\alpha}^{(0)} = \{\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})\}^{-1}\{\sum_{i=1}^n(\mathbf{x}_i-\overline{\mathbf{x}})(Y_i-\overline{Y})\}$, where $\overline{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$ and $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$.
   (b) Write $J(\boldsymbol{\alpha}) = n^{-1/2}\sum_{i=1}^n\{Y_i - \widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}$ and $J'(\boldsymbol{\alpha}) = -n^{-1/2}\sum_{i=1}^n\{\widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}'\{\mathbf{x}_i-\widehat{E}(\mathbf{x}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}\mathbf{x}_i^{\mathrm{T}}$. The derivative $\{\widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}' = \partial\{\widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}/\partial(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)$ is taken directly from the corresponding kernel estimator. Update $\boldsymbol{\alpha}^{(k)}$ with

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} - \{J'(\boldsymbol{\alpha}^{(k)})\}^{-1}\{J(\boldsymbol{\alpha}^{(k)})\}. \tag{2.11}$$

   In case the matrix $J'(\boldsymbol{\alpha}^{(k)})$ is singular or nearly so, we adopt a ridge regression approach using (2.11) with $J'(\boldsymbol{\alpha}^{(k)})$ replaced by $J'_r(\boldsymbol{\alpha}^{(k)}) = J'(\boldsymbol{\alpha}^{(k)}) + \lambda_n\mathbf{I}_{p\times p}$ for some positive ridge parameter λ_n. Here $\mathbf{I}_{p\times p}$ denotes a $p \times p$ identity matrix.
   (c) Iterate (2.11) until $\boldsymbol{\alpha}^{(k+1)}$ fails to change. Denote the resultant estimate by $\widehat{\boldsymbol{\alpha}}$.
2. Obtain $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)$ by using kernel regression of Y_i onto $(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)$, for $i = 1, \ldots, n$.
3. Solve (2.10) to obtain $\widehat{\boldsymbol{\beta}}$, where $\widehat{\varepsilon}_i = Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)$.
   (a) Start with an initial value $\boldsymbol{\beta}^{(0)}$. In our implementation, we choose $\boldsymbol{\beta}^{(0)} = \{\sum_{i=1}^n(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})\}^{-1}\{\sum_{i=1}^n(\mathbf{x}_i - \overline{\mathbf{x}})\widehat{\varepsilon}_i^2\}$.
   (b) Write $I(\boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^n\{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}$ and $I'(\boldsymbol{\beta}) = -n^{-1/2}\sum_{i=1}^n\{\widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}'\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\mathbf{x}_i^{\mathrm{T}}$, where the derivative $\{\widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}' = \partial\{\widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}/\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ is taken from the kernel estimator $\widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$.
   Update $\boldsymbol{\beta}^{(k)}$ with

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \{I'(\boldsymbol{\beta}^{(k)}) + \lambda_n\mathbf{I}_{p\times p}\}^{-1}\{I(\boldsymbol{\beta}^{(k)})\}. \tag{2.12}$$

   (c) Iterate (2.12) until $\boldsymbol{\beta}^{(k+1)}$ fails to change. Denote the resultant estimate by $\widehat{\boldsymbol{\beta}}$.
4. Obtain $\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$ by using kernel regression of $\widehat{\varepsilon}_i^2$ onto $(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$, for $i = 1, \ldots, n$.

One question arises naturally: how to quantify the differences between the estimators obtained from (2.8) and (2.10)? This is answered by the following.

Theorem 1. *Suppose the conditions* (C1)−(C5) *in the Appendix are satisfied. Let*

$$\mathbf{Q} = E\left[\mathbf{A}(\mathbf{x})\left\{\sigma_0^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}'\mathbf{x}^{\mathrm{T}}\right] \quad \textit{and} \quad \mathbf{V} = E\left\{\mathrm{var}(\varepsilon^2 \mid \mathbf{x})\mathbf{A}(\mathbf{x})\mathbf{A}^{\mathrm{T}}(\mathbf{x})\right\},$$

where $\mathbf{A}(\mathbf{x}) = \mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$. *Then* $n^{1/2}\mathbf{V}^{-1/2}\ \mathbf{Q}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ *converges in distribution to the standard multivariate normal distribution as* $n \to \infty$.

Theorem 1 provides the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ obtained from Step 3 in our proposed algorithm, which uses the residual $\widehat{\varepsilon}_i$ and relies on (2.10). As detailed in the proof in the Appendix, we obtain exactly the same asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ based on (2.8) with known error ε_i. This means our procedure performs as well as the oracle procedure in terms of estimating $\boldsymbol{\beta}_0$, namely, without knowing the conditional mean, we can estimate the conditional variance function asymptotically as well as if the conditional mean was given a priori.

With a consistent estimator $\widehat{\boldsymbol{\beta}}$, we estimate $\sigma^2(\cdot)$ via kernel regression. The next theorem states the consistency of $\widehat{\sigma}^2(\cdot)$ obtained from Step 4 in our proposed algorithm.

Theorem 2. *Suppose the conditions* (C1)−(C5) *in the Appendix are satisfied. Let*

$$\mathrm{bias} = h_2^2\mu_2\left[\frac{\left\{\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}''}{2} + \frac{\left\{\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}'\left\{f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}'}{f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})}\right],$$

where $\mu_2 = \int_{-1}^{1} u^2K(u)du$, $\left\{\sigma^2(\cdot)\right\}'$, $\left\{f(\cdot)\right\}'$ *and* $\left\{\sigma^2(\cdot)\right\}''$ *denote the first and second order derivatives, respectively. Then* $(nh_2)^{1/2}\left\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}) - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) - \mathrm{bias}\right\}$ *converges in distribution to normal distribution with mean zero and variance* $\mathrm{var}(\varepsilon^2 \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})/f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$.

Because $\widehat{\boldsymbol{\beta}}$ has a faster convergence rate than the nonparametric regression, the above result is not surprising. Theorem 2 implies that we can estimate $\sigma^2(\cdot)$ based on $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ as efficiently as if $\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ was known a priori. This extends the results in Fan and Yao (1998), as we obtain the adaptive property when $\mathbf{x}$ is high-dimensional and the link function is an unknown smoothing function.

Remark 1. In this section, we describe how to estimate $\boldsymbol{\beta}_0$ from the estimating equation (2.10) and establish the adaptive property for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$ in an asymptotic sense. In practical implementation, one may choose the bandwidth

for the proposed procedure by using a cross-validation procedure. The estimating equation (2.10) corresponds to (2.3) at the population level. Similarly, one can derives an estimate for $\boldsymbol{\beta}_0$ using the sample version of (2.5) or (2.7). However, it is necessary to undersmooth $\hat{E}(\mathbf{x}|\boldsymbol{\beta}_0^T\mathbf{x})$ in (2.5), or $\hat{\sigma}^2(\boldsymbol{\beta}_0^T\mathbf{x})$ in (2.7), in order for the resulting estimate to achieve the adaptive properties. We skip the details.

3. Numerical Studies

3.1. Simulations

In this section, we report on simulation studies to compare the performance of the estimation procedures discussed in Section 2.1. Specifically, we consider

$$\begin{cases} \sum_{i=1}^n \{Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}\{\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}'\mathbf{x}_i = \mathbf{0}, \\ \sum_{i=1}^n \{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}'\mathbf{x}_i = \mathbf{0}. \end{cases} \tag{3.1}$$

Solving (3.1) yields the classical nonlinear least squares estimation proposed by Härdle, Hall, and Ichimura (1993), and it serves as a benchmark for our comparisons.

Estimating equations for the sample level of (2.7) are

$$\begin{cases} \sum_{i=1}^n \{Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}\mathbf{x}_i = \mathbf{0}, \\ \sum_{i=1}^n \{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}\mathbf{x}_i = \mathbf{0}. \end{cases} \tag{3.2}$$

We recall that the quantity $\widetilde{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ is misspecified to be $\mathbf{0}$ in (2.7). We have seen in Section 2.1 that the estimator of $\boldsymbol{\beta}_0$ obtained from (3.2) remains consistent if $\widehat{\varepsilon}^2$ and $\widehat{\sigma}^2(\cdot)$ are consistent. The estimator based on (3.2) is included to demonstrate this robustness property.

The sample version of the estimating equation (2.3) is

$$\begin{cases} \sum_{i=1}^n \{Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}, \\ \sum_{i=1}^n \{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\} = \mathbf{0}. \end{cases} \tag{3.3}$$

In equations (3.1), (3.2), and (3.3), $\widehat{\varepsilon}_i = Y_i - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)$. As in Section 2.2, we estimate $\ell(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})$, $\sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, $E(\mathbf{x} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})$, and $E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ through the corresponding kernel estimates. Both $\sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ are estimated consistently in (3.3). The Epanechnikov kernel is used in our numerical study and the bandwidths are selected via cross-validation. We remark here that, our estimating procedure is not very sensitive to the choice of the bandwidth, which confirms the theoretical investigations in Theorem 1 that the asymptotic normality of $\widehat{\boldsymbol{\beta}}$ holds true for a wide range of bandwidth.

In our simulation, we used two schemes to generate $\mathbf{x}$.

Table 1. Simulation results for Case 1. The oracle estimates use the true error term ε. All numbers reported are multiplied by 100.

method		mean estimate				variance estimate			oracle estimation		
		$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_8$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_8$
(a): $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 5\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$											
(3.1)	bias	-0.04	0.03	0.06	-0.06	7.68	0.07	-13.56	6.04	-0.19	-14.22
	std	0.92	1.19	0.99	1.14	17.14	19.84	11.89	16.78	19.61	12.14
(3.2)	bias	-0.04	0.04	0.05	-0.06	2.62	0.62	-4.16	1.46	0.07	-4.49
	std	0.91	1.18	0.98	1.13	9.67	11.45	5.88	9.06	11.14	5.70
(3.3)	bias	-0.05	0.04	0.05	-0.06	2.48	0.59	-4.04	1.35	0.01	-4.37
	std	0.91	1.18	0.98	1.13	9.35	11.19	5.30	9.01	10.96	5.23
(b): $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 2\exp\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$											
(3.1)	bias	-0.56	-0.16	0.58	-0.49	7.65	-0.29	-13.59	6.04	-0.19	-14.22
	std	3.71	4.82	4.05	4.63	16.84	19.77	11.99	16.78	19.61	12.14
(3.2)	bias	-0.79	-0.11	0.61	-0.45	2.91	0.81	-4.02	1.46	0.07	-4.49
	std	4.06	5.25	4.42	5.07	9.52	11.43	5.59	9.06	11.14	5.70
(3.3)	bias	-0.79	-0.11	0.60	-0.46	2.84	0.79	-3.97	1.35	0.01	-4.37
	std	4.08	5.26	4.41	5.07	9.44	11.25	5.70	9.01	10.96	5.23

Case 1: The predictors $\mathbf{x}$ were drawn from a normal population with mean zero and variance-covariance matrix $(\sigma_{ij})_{8\times 8}$, where $\sigma_{ij} = 0.5^{|i-j|}$.

Case 2: We generated X_1 as uniform $U(0, 12^{1/2})$, X_2 from a binomial distribution with success probability 0.5, and X_3 from a Poisson distribution with parameter 2. We kept $(X_4, \ldots, X_8)^{\mathrm{T}}$ generated from Case 1.

Conditioning on $\mathbf{x}$, Y was generated as normal with the mean functions

(a) $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 5\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$,

(b) $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 2\exp\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$,

where $\boldsymbol{\alpha}_0 = (0.8, 0.4, -0.4, 0.2, 0, 0, 0, 0)^{\mathrm{T}}$ in (a) and (b) and the variance function $\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = 0.25\left(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} + 2\right)^2$ with $\boldsymbol{\beta}_0 = (-0.45, 0, \ldots, 0, 0.9)^{\mathrm{T}}$.

Performances of estimating $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$. For the estimation accuracy of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$, simulations were repeated 1,000 times with sample size $n = 600$. The bias ("bias") and the standard deviation ("std") of the estimates of typical elements of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ are reported in Tables 1 and 2 for Cases 1 and 2, respectively.

The estimating equations (3.1), (3.2), and (3.3) have similar performances for estimating $\widehat{\boldsymbol{\alpha}}$ in terms of both the bias and the standard deviation. The estimating equations (3.3) perform the best for estimating $\boldsymbol{\beta}_0$, and (3.2) perform only slightly worse, confirming the robustness property of (2.2). The estimating equations (3.1) have the largest biases and standard deviations; this may be

Table 2. Simulation results for Case 2. The oracle estimates use the true error term ε. All numbers reported are multiplied by 100.

method		mean estimate				variance estimate			oracle estimation		
		$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_8$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_8$
(a): $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 5\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$											
(3.1)	bias	-0.02	0.00	0.01	-0.03	5.98	1.65	-13.35	5.68	1.27	-13.74
	std	0.49	1.09	0.16	0.72	15.56	28.91	12.50	16.10	29.08	12.87
(3.2)	bias	-0.02	-0.01	0.00	-0.03	-0.05	-2.40	-5.39	-0.45	-3.07	-5.33
	std	0.50	1.12	0.42	0.72	8.69	17.05	6.73	7.95	16.86	6.13
(3.3)	bias	-0.02	-0.00	0.01	-0.03	1.86	0.26	-3.75	1.66	-0.44	-3.67
	std	0.50	1.12	0.42	0.73	7.61	16.64	4.83	7.41	16.36	4.90
(b): $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = 2\exp\left(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}\right)$											
(3.1)	bias	1.07	1.42	4.28	-0.91	5.42	2.25	-13.99	5.68	1.27	-13.74
	std	2.08	4.06	2.89	2.50	16.17	28.87	13.07	16.10	29.08	12.87
(3.2)	bias	-0.86	-1.43	-1.69	0.17	0.21	-2.63	-5.43	-0.45	-3.07	-5.33
	std	2.24	5.02	2.23	3.21	9.03	17.52	7.03	7.95	16.86	6.13
(3.3)	bias	-0.39	-0.21	0.17	-0.18	2.02	0.36	-3.66	1.66	-0.44	-3.67
	std	2.22	4.89	1.99	3.15	7.46	16.73	4.90	7.41	16.36	4.90

Table 3. Simulation results of the average squared errors (ASE). All numbers reported are multiplied by 100.

model	method	Case 1		Case 2	
		$ASE(\widehat{\boldsymbol{\alpha}})$	$ASE(\widehat{\boldsymbol{\beta}})$	$ASE(\widehat{\boldsymbol{\alpha}})$	$ASE(\widehat{\boldsymbol{\beta}})$
(a)	(3.1)	3.83	98.56	1.84	41.34
	(3.2)	3.83	43.64	1.94	18.47
	(3.3)	3.84	43.24	1.94	17.89
(b)	(3.1)	4.30	96.86	3.05	42.23
	(3.2)	4.25	45.10	2.39	18.59
	(3.3)	4.25	44.35	2.21	18.26

explained by the fact that estimation of $\{\ell(\cdot)\}'$ and $\{\sigma^2(\cdot)\}'$ are involved. From Tables 1 and 2, all three procedures are close to their corresponding oracle estimates, which adopt the true errors instead of the residuals. This serves to confirm the adaptive property of the proposed estimators.

Performances of estimating $\ell(\cdot)$ and $\sigma^2(\cdot)$. We take the average squared errors criteria

$$ASE(\widehat{\boldsymbol{\alpha}}) = n^{-1}\sum_{i=1}^{n}\left\{\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i) - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}^2,$$

$$ASE(\widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n}\left\{\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i) - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\right\}^2.$$

The results are reported in Table 3. The estimating equations (3.2) and (3.3) offer similar results, and both are superior to the results based on (3.1).

3.2. An application

We consider the horse mussels data collected in the Malborough Sounds at the Northeast of New Zealands South Island (Camden (1989)). The response variable Y is the shell mass, with three quantitative predictors as related characteristics of mussel shells: the shell length X_1, the shell height X_2, and the shell width X_3, all in mm. The sample size of this data set is 201. All predictors in this analysis are standardized marginally to have zero mean and unit variance.

We only report the results obtained from (3.3), as similar results are obtained from (3.1) and (3.2). By assuming (1.2), we estimate $\boldsymbol{\alpha}_0$ by the first equation of (3.3), and find $\widehat{\boldsymbol{\alpha}} = (0.3615, 0.4243, 0.8302)^{\mathrm{T}}$. Based on $\{(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$, we estimate $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$ by kernel regression. The estimated regression function and its point-wise prediction interval are plotted in Figure 1(A). The prediction interval at the 95% level is calculated by assuming tentatively that the variance function is a constant. We can clearly see that the empirical converge probability is very poor, particularly when $\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}$ is large, only 63 among 201 points lie within this prediction region. Homoscedasticity is not a reasonable assumption for this data set.

Next we solve the second of the estimating equations (3.3) and find $\widehat{\boldsymbol{\beta}} = (-0.3792, 0.3724, 0.8471)^{\mathrm{T}}$. The variance function is then estimated by kernel regression based on data points $\{(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i, \widehat{\varepsilon}_i^2), i = 1, \ldots, n\}$. The estimated variance function is plotted in Figure 1(B), which does not seem to be constant. Taking into account the heteroscedasticity, we report the 95% prediction interval of $\widehat{\ell}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$ in Figure 1(C). We can see that, around 94% of the sample (189 points) is covered by this prediction interval.

It is of practical interest to examine the adaptive property of our proposal by using the bootstrap method. Based on the original sample, we obtained the estimates of the index parameters $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ and the link function $\widehat{\ell}(\cdot)$. We then bootstrapped the original data 1,000 times. Three quantities can be obtained from the bootstrap sample: $\widehat{\boldsymbol{\alpha}}^*$ denotes the estimator of $\boldsymbol{\alpha}_0$ based on the bootstrap sample $(\mathbf{x}_i^*, Y_i^*)$; $\widehat{\boldsymbol{\beta}}^*$ denotes the estimator of $\boldsymbol{\beta}_0$ based on $\{\mathbf{x}_i^*, Y_i^* - \widehat{\ell}^*(\mathbf{x}_i^{*\mathrm{T}}\widehat{\boldsymbol{\alpha}}^*)\}$, and $\widehat{\boldsymbol{\beta}}_o^*$ denotes the estimator of $\boldsymbol{\beta}_0$ based on $\{\mathbf{x}_i^*, Y_i^* - \widehat{\ell}(\mathbf{x}_i^{*\mathrm{T}}\widehat{\boldsymbol{\alpha}})\}$. We remark here that $\widehat{\boldsymbol{\beta}}_o^*$ differs from $\widehat{\boldsymbol{\beta}}^*$ in that the former used the "true error" term because it adopted $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\ell}(\cdot)$, estimated from the original data, while the latter used the residuals calculated from the bootstrap sample. The third plot in Figure 1(D) gives the boxplot of the absolute values of correlation coefficients $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_0)$. It implies that $\widehat{\boldsymbol{\beta}}^*$ behaves very similarly to $\widehat{\boldsymbol{\beta}}_0^*$ because the correlation coefficients

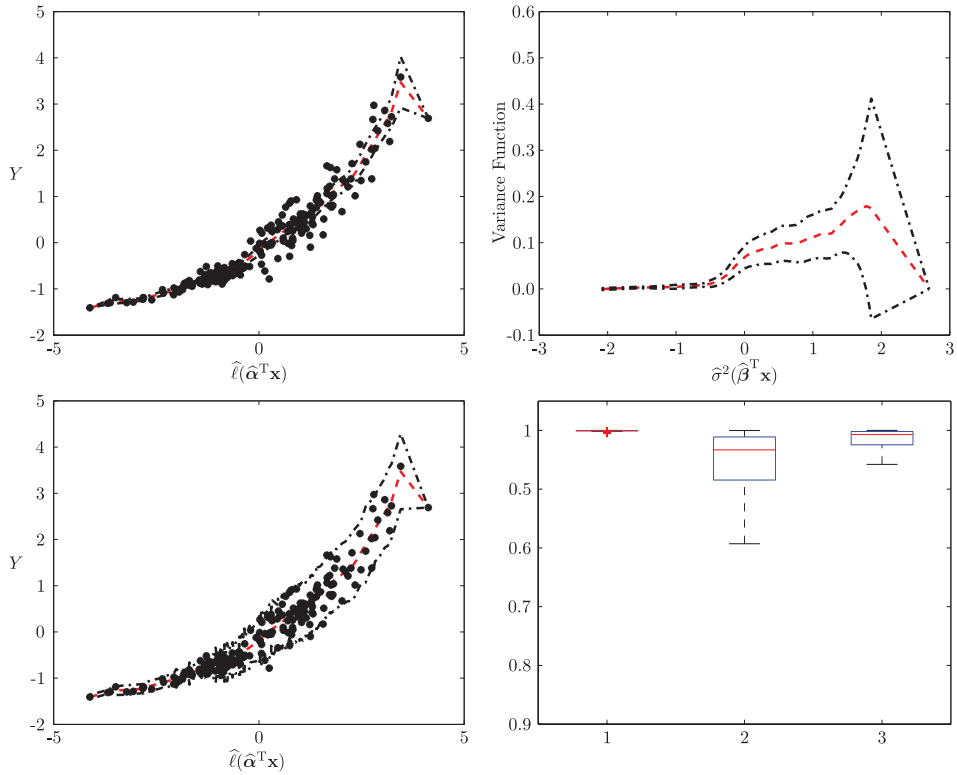

Figure 1. Analysis of the Horse Mussel Data. The dashed line in (A) is the kernel estimate $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})$. The dot-dash lines are the 95% pointwise confidence intervals obtained under the homoscedasticity assumption. The dashed line in (B) is the kernel estimate $\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$. The dashed line in (C) is the kernel estimate $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})$. The dot-dash lines are the 95% pointwise confidence intervals obtained under the heteroscedasticity assumption. (D) depicts the boxplots of $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}})$, $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$, and $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_0)$, as calculated from the bootstrap samples.

are all very large. The first and second plots in Figure 1(D) show the respective boxplots of the absolute value of the correlation coefficients $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}})$ and $\mathrm{corr}(\mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^*, \mathbf{x}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$. It can be seen that $\widehat{\boldsymbol{\alpha}}^*$ performs much more stably than $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\beta}}_0^*$, indicating that estimating the conditional variance function is more difficult than estimating the conditional mean function.

4. Discussion

Estimation of the conditional heteroscedasticity remains an important and open problem in the literature when the dimension of predictors is very large (Antoniadis, Grégoire, and Mckeague (2004)). Based on a single-index structure,

this paper offers a general class of estimating equations to estimate the conditional heteroscedasticity. The proposed framework allows for flexibility when the structure of the conditional variance function is unknown. The resulting estimator enjoys an adaptive property in that it performs as well as if the true mean function was given. For ease of illustration, we assume that both the conditional mean and the conditional variance functions are of the single-index structure. Extension of the proposed methodology to the multi-index conditional mean and conditional variance models warrants future investigation. As pointed out by an anonymous referee, the model considered for ε is of the form $\varepsilon = \sigma(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\epsilon$. Thus, one can estimate $\boldsymbol{\beta}_0$ by considering $|\varepsilon|^\alpha = \sigma^\alpha(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})E|\epsilon|^\alpha + \xi$ for any $\alpha > 0$, where $\xi = \sigma^\alpha(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})(|\epsilon|^\alpha - E|\epsilon|^\alpha)$. The model used in this paper corresponds to $\alpha = 2$. Mercurio and Spokony (2004) discussed how to determine α, and suggested $\alpha = 1/2$ in general. It may be of interest to consider other α than $\alpha = 2$. This is beyond the scope of this paper.

Acknowledgement

Zhu's research was supported by National Natural Science Foundation of China (NNSFC) 11071077, the National Institute on Drug Abuse (NIDA) grant R21-DA024260, Innovation Program of Shanghai Municipal Education Commission 13ZZ055 and Pujiang Project of Science and Technology Commission of Shanghai Municipality 12PJ1403200. Dong's research was supported by National Science Foundation Grant DMS-1106577. Runze Li is the corresponding author and his research was supported in part by NNSFC grants 11028103 and 10911120395, NIDA grant P50-DA10075 and NCI grant R01 CA168676. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA, NCI or the National Institutes of Health.

Appendix: Technical Conditions and Proofs

Appendix A: Some technical conditions

(C1) The density functions of $(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$ and $(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, denoted by $f_{\boldsymbol{\alpha}_0}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$ and $f_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, are continuous and bounded from zero and above for all $\mathbf{x} \in \mathbb{R}^p$, and have locally Lipschitz second derivatives.

(C2) The functions $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$, $\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, $\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})f_{\boldsymbol{\alpha}_0}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$, $\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})f_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, $E(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})f_{\boldsymbol{\alpha}_0}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$, and $E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})f_{\boldsymbol{\beta}_0}(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ are continuous and bounded from above for all $\mathbf{x} \in \mathbb{R}^p$, and their third derivatives are locally Lipschitz.

(C3) The symmetric kernel function $K(\cdot)$ is twice continuously differentiable with support [-1,1], and is Lipschitz continuous.

(C4) The bandwidths h, h_1, and h_2 satisfy $nh^4 \to \infty$ and $nh^8 \to 0$, $nh_i^4 \to \infty$ and $nh_i^8 \to 0$ for $i = 1$ and 2.

(C5) $E(Y^4) < \infty$ and $E(X_i^4) < \infty$ for $i = 1, \ldots, p$. In addition, the conditional variance of Y given $\mathbf{x}$ is bounded away from 0 and from above.

Appendix B: Proofs of Theorems 1 and 2

Proof of Theorem 1. With regard to $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})$, $\widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$, and $\widehat{E}(\mathbf{x} \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$, we first quantify the extent to which these functions can approximate their respective true values. We take $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x})$ as an example. Note that

$$\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) = \left\{\widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\} + \left\{\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}.$$

The first term in the right hand side can be dealt with using standard nonparametric regression theory. Therefore, it remains to quantify the difference $\widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$. Let $\mathcal{C} = \{\widetilde{\boldsymbol{\alpha}} : \|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| \le Cn^{-1/2}\}$. Recall that Li, Zhu, and Zhu (2011, Lemma 1) proved that, if $nh_1^4 \to \infty$,

$$\sup_{\mathbf{x}\in\mathbb{R}^p} \sup_{\mathcal{C}} \left| \left\{\widehat{\ell}(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\} - E\left\{\widehat{\ell}(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\} \right| = O_p\left\{\frac{\log n}{nh_1^2}\right\}.$$

By invoking the symmetry of the kernel function and the condition that the third derivative of $\ell(\cdot)f(\cdot)$ is local Lipschitz, and using similar arguments as those in proving Lemma 3.3 of Zhu and Fang (1996), we can show that

$$E\left\{\widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\} - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \mu_2 h_1^2 \left\{\ell''(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) + \frac{\ell'(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})f'(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})}{f(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})}\right\} = O(h^4),$$

where $\mu_2 = \int_{-1}^{1} u^2 K(u)du$. A similar result also holds when $\boldsymbol{\alpha}_0$ is replaced by $\widetilde{\boldsymbol{\alpha}}$. By the Mean Value Theorem, it follows that

$$\sup_{\mathcal{C}} |\ell''(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) - \ell''(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})| h_1^2 \mu_2 = O_p\left(\frac{h_1^2}{n^{1/2}}\right) = O(h_1^4),$$

since $nh_1^4 \to \infty$. The above arguments imply that, for any fixed $\mathbf{x} \in \mathbb{R}^p$ and $nh_1^8 \to 0$,

$$\widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}) = \ell'(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\mathbf{x}^{\mathrm{T}}(\boldsymbol{\alpha}_0 - \widetilde{\boldsymbol{\alpha}}) + o_p(n^{-1/2}). \tag{B.1}$$

Following similar arguments, we can obtain that

$$\widehat{\sigma}_0^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\sigma}_0^2(\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}) = \left\{\sigma_0^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}' \mathbf{x}^{\mathrm{T}}\left(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right) + o_p(n^{-1/2}),$$

and

$$\widehat{E}(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) - \widehat{E}(\mathbf{x} \mid \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}) = \left\{E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}' \mathbf{x}^{\mathrm{T}} \left(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}\right) + o_p(n^{-1/2}). \quad \text{(B.2)}$$

(B.1) and (B.2) are used repetitively in subsequent derivations. We discuss the consistency of $\widehat{\boldsymbol{\beta}}$ first. Let

$$I(\boldsymbol{\beta}) = E\left\{\varepsilon^2 - \sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\},$$
$$\widehat{I}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} E\left\{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}.$$

Let $U(\boldsymbol{\beta}_0)$ be any open set that includes $\boldsymbol{\beta}_0$. To prove the consistency of $\widehat{\boldsymbol{\beta}}$, we assume that $\inf_{\boldsymbol{\beta}\in\Theta\setminus U(\boldsymbol{\beta}_0)} \mid I(\boldsymbol{\beta}) \mid \geq \delta$ for some positive constant δ. It suffices to show that

$$Pr\Big\{\inf_{\boldsymbol{\beta}\in\Theta\setminus U(\boldsymbol{\beta}_0)} \mid \widehat{I}(\boldsymbol{\beta}) \mid \leq \frac{\delta}{2}\Big\} \to 0. \quad \text{(B.3)}$$

The condition $\inf_{\boldsymbol{\beta}\in\Theta\setminus U(\boldsymbol{\beta}_0)} \mid I(\boldsymbol{\beta}) \mid \geq \delta$ implies that

$$Pr\Big\{\inf_{\boldsymbol{\beta}\in\Theta\setminus U(\boldsymbol{\beta}_0)} \mid \widehat{I}(\boldsymbol{\beta}) \mid \leq \frac{\delta}{2}\Big\} \leq Pr\Big\{\inf_{\boldsymbol{\beta}\in\Theta\setminus U(\boldsymbol{\beta}_0)} \mid \widehat{I}(\boldsymbol{\beta}) - I(\boldsymbol{\beta}) \mid \geq \frac{\delta}{2}\Big\}.$$

Therefore, it suffices to show that, for any fixed δ,

$$Pr\Big\{\sup_{\boldsymbol{\beta}\in\Theta} \mid \widehat{I}(\boldsymbol{\beta}) - I(\boldsymbol{\beta}) \mid \geq \frac{\delta}{2}\Big\} \to 0.$$

Recall the definition of $\widehat{I}(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$, and let $\widehat{I}(\boldsymbol{\beta}) - I(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \sum_{i=1}^{6} I_{1,i}$, where

$$I_{1,1} = n^{-1}\sum_{i=1}^{n}\left\{\varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\} - I(\boldsymbol{\beta}),$$
$$I_{1,2} = n^{-1}\sum_{i=1}^{n}\left\{\widehat{\varepsilon}_i^2 - \varepsilon_i^2\right\}\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\},$$
$$I_{1,3} = n^{-1}\sum_{i=1}^{n}\left\{\sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\},$$
$$I_{1,4} = n^{-1}\sum_{i=1}^{n}\left\{\varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\},$$
$$I_{1,5} = n^{-1}\sum_{i=1}^{n}\left\{\widehat{\varepsilon}_i^2 - \varepsilon_i^2\right\}\left\{E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\},$$
$$I_{1,6} = n^{-1}\sum_{i=1}^{n}\left\{\sigma^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}.$$

We note that $I_{1,1}$ is an average of independent and identically distributed random variables with mean zero. The classical theory of empirical process shows that $\sup_{\boldsymbol{\beta}\in\Theta} \mid I_{1,1} \mid = o_p(\log n/n^{1/2})$ almost surely (Pollard (1984, Thm. 2.37)). Write

$$I_{1,2} = 2n^{-1}\sum_{i=1}^{n}\left[\left\{\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\right\}\varepsilon_i\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\right]$$
$$+n^{-1}\sum_{i=1}^{n}\left[\left\{\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\right\}^2\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\right] = 2I_{1,2,1} + I_{1,2,2}.$$

We first prove that $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,2,1}| = o_p(1)$. Note that

$$\begin{aligned}|I_{1,2,1}| &\le \sup_{\mathbf{x}\in\mathbb{R}^p}\left|\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right| n^{-1}\sum_{i=1}^{n}\left|\varepsilon_i\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\right| \\ &\quad +n^{-1}\sum_{i=1}^{n}\left|\left\{\widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{\ell}(\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{x}_i)\right\}\right|\left|\varepsilon_i\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\right| \\ &\le \sup_{\mathbf{x}\in\mathbb{R}^p}\left|\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right| n^{-1}\sum_{i=1}^{n}\left|\varepsilon_i\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\right| \\ &\quad +2n^{-1}\sum_{i=1}^{n}\left|\varepsilon_i\ell'(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right\}\mathbf{x}_i^{\mathrm{T}}\right|\left|\boldsymbol{\alpha}_0 - \widehat{\boldsymbol{\alpha}}\right|.\end{aligned}$$

Therefore, $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,2,1}| = o_p(1)$ follows immediately from the consistency of $\widehat{\boldsymbol{\alpha}}$ and the uniform consistency of $\widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})$. Similarly, $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,2,2}| = o_p(1)$ can be proved. These two results imply that $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,2}| = o_p(1)$. Using the U-statistic theory (Serfling (1980)), we can obtain that $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,i}| = o_p(1), i = 3, 4$. Following similar arguments to deal with $I_{1,2}$, we have $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,5}| = o_p(1)$. That $\sup_{\boldsymbol{\beta}\in\Theta}|I_{1,6}| = o_p(1)$ follows directly from the Cauchy-Schwartz inequality and standard results on the uniform convergence of $\widehat{E}(\mathbf{x} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\widehat{\sigma}^2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ in nonparametric regression. By combining these results, it follows that

$$Pr\Big\{\sup_{\boldsymbol{\beta}\in\Theta} \mid \widehat{I}(\boldsymbol{\beta}) - I(\boldsymbol{\beta}) \mid \ge \frac{\delta}{2}\Big\} \to 0$$

for any fixed δ, which completes the proof of (B.3).

Next we examine the root-n consistency of $\widehat{\boldsymbol{\beta}}$. We expand the estimating equation as $0 = n^{-1/2}\sum_{i=1}^{n}\left\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\widehat{\varepsilon}_i^2 - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\right\} =: \sum_{i=1}^{12} I_{2,i}$. The terms $I_{2,i}$'s are explicitly defined in the sequel.

By the Central Limit Theorem,

$$I_{2,1} =: n^{-1/2} \sum_{i=1}^{n} \left\{ \mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} = O_p(1).$$

Invoking the root-n consistency of $\widehat{\boldsymbol{\alpha}}$, we can show that

$$I_{2,2} =: n^{-1/2} \sum_{i=1}^{n} \left\{ \mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \widehat{\varepsilon}_i^2 - \varepsilon_i^2 \right\} = o_p(1).$$

Invoking (B.2), it follows that

$$I_{2,3} = \left[E \left\{ \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\}' \left\{ \mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\} \mathbf{x}^{\mathrm{T}} + o_p(1) \right] n^{1/2} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right).$$

Because $E \left\{ \mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i \right\} = 0$, U-statistic theory implies

$$I_{2,4} =: n^{-1/2} \sum_{i=1}^{n} \left\{ \mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) - \widehat{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\} = o_p(1),$$

$$I_{2,5} =: n^{-1/2} \sum_{i=1}^{n} \left\{ E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\} = o_p(1).$$

By the uniform convergence of $\widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i)$, and (B.1), it is easy to see

$$I_{2,6} =: n^{-1/2} \sum_{i=1}^{n} \left\{ E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \widehat{\varepsilon}_i^2 - \varepsilon_i^2 \right\} = o_p(1),$$

$$I_{2,7} =: n^{-1/2} \sum_{i=1}^{n} \left\{ E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \widehat{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}) \right\} = o_p(1).$$

Given $nh^4 h_2^4 \to 0$ and $nhh_2 \to \infty$, the Cauchy-Schwartz inequality implies

$$I_{2,8} =: n^{-1/2} \sum_{i=1}^{n} \left\{ E(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) - \widehat{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\} = 0.$$

By using the fact that $E \left\{ \varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \mid \mathbf{x} \right\} = 0$, we can show that

$$I_{2,9} =: n^{-1/2} \sum_{i=1}^{n} \left\{ \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}) \right\},$$

$$\begin{aligned} I_{2,10} &=: n^{-1/2} \sum_{i=1}^{n} \left\{ \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{x}_i) \right\} \left\{ \widehat{\varepsilon}_i^2 - \varepsilon_i^2 \right\} \\ &= o_p(1) n^{1/2} \left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}} \right). \end{aligned}$$

Invoking (B.2) again, we have

$$\begin{aligned} I_{2,11} &=: n^{-1/2}\sum_{i=1}^{n}\left\{\widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\widehat{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\sigma}^2(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})\right\} \\ &= o_p(1)n^{1/2}\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right), \\ I_{2,12} &=: n^{-1/2}\sum_{i=1}^{n}\left\{\widehat{E}(\mathbf{x}_i \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\sigma}^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\} \\ &= o_p(1)n^{1/2}\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right). \end{aligned}$$

By combining these results, we obtain that $\left(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\right) = O_p(n^{-1/2})$.

The asymptotic normality follows immediately from the root-n consistency described above and the Central Limit Theorem. That is,

$$\begin{aligned} &E\left[\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\sigma_0^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\}'\mathbf{x}^{\mathrm{T}}\right] n^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \\ &\quad = n^{-1/2}\sum_{i=1}^{n}\left\{\mathbf{x}_i - E(\mathbf{x} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\varepsilon_i^2 - \sigma^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})\right\} + o_p(1), \end{aligned}$$

which completes the proof of Theorem 1.

Proof of Theorem 2. It can be proved, following (B.2), the root-n consistency of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, and standard arguments in nonparametric regression. See, for example, Ruppert and Wand (1994). Details are omitted from here.

Appendix C: Some preliminary results

We discuss the consistency of $\widehat{\boldsymbol{\alpha}}$, its convergence rate, and its asymptotic normality. First we prove the consistency of $\widehat{\boldsymbol{\alpha}}$. Take

$$\begin{aligned} J(\boldsymbol{\alpha}) &= E\left\{Y - \ell(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\right\}, \\ \widehat{J}(\boldsymbol{\alpha}) &= n^{-1}\sum_{i=1}^{n}\left\{Y_i - \widehat{\ell}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\right\}. \end{aligned}$$

Let $U(\boldsymbol{\alpha}_0)$ be any open set that includes $\boldsymbol{\alpha}_0$. To prove the consistency of $\widehat{\boldsymbol{\alpha}}$, we assume that $\inf_{\boldsymbol{\alpha}\in\Theta\setminus U(\boldsymbol{\alpha}_0)} \mid J(\boldsymbol{\alpha}) \mid \geq \delta$ for some positive constant δ. It suffices to show

$$Pr\left\{\inf_{\boldsymbol{\alpha}\in\Theta\setminus U(\boldsymbol{\alpha}_0)} \mid \widehat{J}(\boldsymbol{\alpha}) \mid \leq \frac{\delta}{2}\right\} \to 0. \tag{C.1}$$

The condition $\inf_{\boldsymbol{\alpha}\in\Theta\setminus U(\boldsymbol{\alpha}_0)} \mid J(\boldsymbol{\alpha}) \mid \geq \delta$ implies that

$$Pr\left\{\inf_{\boldsymbol{\alpha}\in\Theta\setminus U(\boldsymbol{\alpha}_0)} \mid \widehat{J}(\boldsymbol{\alpha}) \mid \leq \frac{\delta}{2}\right\} \leq Pr\left\{\inf_{\boldsymbol{\alpha}\in\Theta\setminus U(\boldsymbol{\alpha}_0)} \mid \widehat{J}(\boldsymbol{\alpha}) - J(\boldsymbol{\alpha}) \mid \geq \frac{\delta}{2}\right\}.$$

Therefore, it suffices to show that

$$Pr\left\{\sup_{\boldsymbol{\alpha}\in\Theta} \mid \widehat{J}(\boldsymbol{\alpha}) - J(\boldsymbol{\alpha}) \mid \geq \frac{\delta}{2}\right\} \to 0$$

for any fixed δ. This follows directly from the consistency of $\widehat{J}(\boldsymbol{\alpha})$ if $h_1 \to 0$ and $nh_1/\log n \to \infty$. See, for example, Ichimura (1993, Lemmas 5.1-5.3), Zhu and Fang (1996, Lemmas 3.1 and 3.3), and Pollard (1984, Thm. 2.37). (C.1) is proved.

Next we prove the root-n consistency of $\widehat{\boldsymbol{\alpha}}$. With probability close to 1, by a Taylor expansion, for $\widetilde{\boldsymbol{\alpha}}$ between $\boldsymbol{\alpha}_0$ and $\widehat{\boldsymbol{\alpha}}$,

$$\begin{aligned}\mathbf{0} = \widehat{J}(\widehat{\boldsymbol{\alpha}}) &= \widehat{J}(\boldsymbol{\alpha}_0) + \widehat{J}'(\widetilde{\boldsymbol{\alpha}})\,(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \\ &= \widehat{J}(\boldsymbol{\alpha}_0) + \left\{\widehat{J}'(\widetilde{\boldsymbol{\alpha}}) - \widehat{J}'(\boldsymbol{\alpha}_0)\right\}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) + \widehat{J}'(\boldsymbol{\alpha}_0)\,(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)\,. \qquad \text{(C.2)}\end{aligned}$$

Recall the definition of $\widehat{J}(\widehat{\boldsymbol{\alpha}})$ and write $\widehat{J}(\boldsymbol{\alpha}_0) \stackrel{\text{def}}{=} \sum_{i=1}^4 \widehat{J}_i$, where

$$\begin{aligned}\widehat{J}_1 &= n^{-1}\sum_{i=1}^n \left\{Y_i - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}, \\ \widehat{J}_2 &= n^{-1}\sum_{i=1}^n \left\{Y_i - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{E(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}, \\ \widehat{J}_3 &= n^{-1}\sum_{i=1}^n \left\{\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}, \\ \widehat{J}_4 &= n^{-1}\sum_{i=1}^n \left\{\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}\left\{E(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i) - \widehat{E}(\mathbf{x}_i \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)\right\}.\end{aligned}$$

Because $E\left[\left\{Y - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}\right] = 0$, $\widehat{J}_1$ is of the order $O_p(n^{-1/2})$ by the Central Limit Theorem. Both $\widehat{J}_2$ and $\widehat{J}_3$ are of the order $o_p(n^{-1/2})$ according to the standard U-statistics theory (Serfling (1980)); by using Cauchy-Schwartz inequality, $\widehat{J}_4$ is less than

$$\left[\widehat{E}\left\{\ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \widehat{\ell}(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}^2\right]^{1/2}\left[E\left\{E(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}) - \widehat{E}(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}^2\right]^{1/2},$$

which is of order $O_p\left\{h_1^4 + \log n/(nh_1)\right\}$. If $nh_1^8 \to 0$ and $nh_1^2/\log n \to \infty$, then the last term is also $o_p(n^{-1/2})$. Combining these results, we obtained that $\widehat{J}(\boldsymbol{\alpha}_0) = O_p(n^{-1/2})$.

By using the Weak Law of Large Numbers, we can see that $\widehat{J}'(\boldsymbol{\alpha}_0)$ converges in probability to $J'(\boldsymbol{\alpha}_0) = E\left[\ell'(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}\mathbf{x}^{\mathrm{T}}\right]$ for $nh_1^2 \to \infty$ and

$nh_1^8 \to 0$. Similarly, $\widehat{J}'(\widetilde{\boldsymbol{\alpha}}) - \widehat{J}'(\boldsymbol{\alpha}_0) = o_p(1)$ in that $\widehat{\boldsymbol{\alpha}}$ is a consistent estimate of $\boldsymbol{\alpha}_0$. Therefore, $\widehat{\boldsymbol{\alpha}}$ is a root-n consistent estimate of $\boldsymbol{\alpha}_0$.

The asymptotic normality of $\widehat{\boldsymbol{\alpha}}$ follows from its root-n consistency. Specifically, (C.2) implies that

$$n^{1/2}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\right) = \left\{J'(\boldsymbol{\alpha}_0)\right\}^{-} n^{-1/2} \sum_{i=1}^{n} \varepsilon_i \left\{\mathbf{x}_i - E(\mathbf{x}_i \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\right\} + o_p(1)$$

where $J'(\boldsymbol{\alpha}_0) = E\left[\ell'(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\left\{\mathbf{x} - E(\mathbf{x} \mid \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x})\right\}\mathbf{x}^{\mathrm{T}}\right]$ and $\varepsilon_i = Y_i - \ell(\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}_i)$.

References

Anderson, T. G. and Lund, J. (1997). Estimating continuous time stochastic volatility models of the short term interest rate. *J. Econometrics* **77**, 343-377.

Antoniadis, A., Grégoire, G. and Mckeague, I. (2004). Bayesian estimation in single-index models. *Statist. Sinica* **14**, 1147-1164.

Box, G. E. P and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc. Ser. B* **26**, 211-252.

Cai, T., Levine, M. and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.* **100**, 126-136.

Camden, M. (1989). *The Data Bundle*. New Zealand Statistical Association, WeHington, New Zealand.

Carroll, R. J. and Ruppert, D. (1988). *Transformations and Weighting in Regression*. Chapman & Hall, London.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

Hall, P. and Carroll, R. J. (1989). Variance function estimating in regression: the effect of estimating the mean. *J. Roy. Statist. Soc. Ser. B* **51**, 3-14.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58**, 71-120.

Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**, 1272-1298.

Li, K.-C. (1991), Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, L. X., Zhu, L. P. and Zhu, L. X. (2011). Inference on primary parameter of interest with aid of dimension reduction estimation. *J. Roy. Statist. Soc. Ser. B* **73**, 59-80.

Mercurio, D. and Spokoiny, V. (2004). Statistical inference for time-inhomogeneous volatility models. *Ann. Statist.* **32**, 577-602

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610-625.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

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

Ruppert, D., Wand, M., Holst, U. and Hössjer, O (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262-273.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

Song, Q. and Yang, L. (2009). Spline confidence bands for variance function. *J. Nonparametr. Statist.* **21**, 589-609.

Wang, L., Brown, L. D., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation on nonparametric regression. *Ann. Statist.* **36**, 646-664.

Xia, Y., Tong, H. and Li, W. K. (2002). Single-index volatility models and estimation. *Statist. Sinica* **12**, 785-799.

Yao, Q. and Tong, H. (1994). Quantifying the influence of initial values on nonlinear prediction. *J. Roy. Statist. Soc. Ser. B* **56**, 701-725.

Yin, J., Geng, Z., Li, R. and Wang, H. (2010). Nonparametric covariance model. *Statist. Sinica* **20**, 469-479.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053-1068.

School of Statistics and Management and the Key Laboratory of Mathematical Economics, Ministry of Education, Shanghai University of Finance and Economics, Shanghai 200433, P. R. China.

E-mail: zhu.liping@mail.shufe.edu.cn

Department of Statistics, Temple University, Philadelphia, PA 19122, U. S. A.

E-mail: ydong@temple.edu

Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA 16802-2111, U.S.A.

E-mail: rli@stat.psu.edu

(Received March 2012; accepted August 2012)