Published in final edited form as:

Stat Sin. 2013 July ; 23(3): 1215–1235. doi:10.5705/ss.2012.075.

SEMIPARAMETRIC ESTIMATION OF CONDITIONAL HETEROSCEDASTICITY VIA SINGLE-INDEX MODELING

Liping Zhu,

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, zhu.liping@mail.shufe.edu.cn

Yuexiao Dong, and

Department of Statistics, Temple University, Philadelphia, PA 19122, U. S. A., ydong@temple.edu

Runze Li

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

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 also possess an adaptive property in an asymptotic sense. That is, they 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 estimator is superior to existing estimators with less bias and smaller standard deviation.

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, which is 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 et al., 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 thus an important problem in statistics.

Let $Y \in \mathbb{R}$ be the response variable and $\mathbf{x} = (X_1, \dots, X_p)^T \in \mathbb{R}^p$ be the associated predictor vector. In this paper we study the conditional variance function of Y given $\mathbf{x}$, denoted as $\text{var}(Y | \mathbf{x})$. Denote $E(Y | \mathbf{x})$ as the conditional mean function and $\varepsilon = Y - E(Y | \mathbf{x})$ as the random error. Then we have $\text{var}(Y | \mathbf{x}) = E(\varepsilon^2 | \mathbf{x})$. Since ε is not observable, it is natural to replace the error with the residual $\varepsilon = Y - \hat{E}(Y | \mathbf{x})$, where $\hat{E}(Y | \mathbf{x})$ 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 has first received

much attention when the predictor is univariate with $p = 1$. See, for example, Hall and Carroll (1989), Wang et al. (2008) and references therein. Cai et al. (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 was 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 this paper that there exists a smooth function $\sigma^2(\cdot)$ and a $p \times 1$ vector β_0 such that

$$\text{var}(Y|\mathbf{x}) = \sigma^2(\beta_0^T \mathbf{x}). \quad (1.1)$$

With an unspecified link function $\sigma^2(\cdot)$ and a univariate index $\beta_0^T \mathbf{x}$, model (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 the ease of presentation, we assume that for some unknown function $\ell(\cdot)$ and a $p \times 1$ vector α_0 , the conditional mean function also admits the single-index structure

$$E(Y|\mathbf{x}) = \ell(\alpha_0^T \mathbf{x}). \quad (1.2)$$

Different from model (1.1), the assumption of model (1.2) is not essential and more general forms of the mean function may be assumed.

In this paper we make the following contributions to the literature.

1. We propose a general class of estimating equations to estimate β_0 in this paper. By correctly specifying $E(\mathbf{x}|\beta_0^T \mathbf{x})$, the estimator of β_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 β_0 will be consistent without correctly specifying $E(\mathbf{x}|\beta_0^T \mathbf{x})$.
2. The resulting estimate of β_0 from the aforementioned estimating equations possesses an adaptive property asymptotically. That is, the proposed procedure estimates the conditional variance function as efficiently as if the conditional mean was given a priori asymptotically. This adaptive property is achieved by replacing the error $\varepsilon = Y - E(Y|\mathbf{x})$ in the estimating equation with their corresponding residual $\hat{\varepsilon} = Y - \hat{E}(Y|\mathbf{x})$.

This paper is organized in the following way. In Section 2 we present the methodology and study its theoretical 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 real dataset. This paper is concluded 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

We first consider the following estimating equation,

$$E[\{\varepsilon^2 - \sigma^2(\beta_0^T \mathbf{x})\} \{\sigma^2(\beta_0^T \mathbf{x})\}' | \mathbf{x}] = \mathbf{0}, \quad (2.1)$$

where $\varepsilon = Y - E(Y|\mathbf{x})$, and $\{\sigma^2(\cdot)\}'$ stands for the first order derivative of $\sigma^2(\cdot)$. The estimating equation (2.1) corresponds to the classical nonlinear least squares estimation proposed by Ichimura (1993) and Härdle et al. (1993). A limitation of equation (2.1) is that it requires correct specification of $\sigma^2(\cdot)$ to obtain a consistent estimate for β_0 , which may be difficult in practice. To address this issue we propose a new class of estimating equations,

$$E[\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x})\} \{\mathbf{x} - \tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})\}] = \mathbf{0}, \quad (2.2)$$

where $\tilde{\sigma}^2(\beta_0^T \mathbf{x})$ and $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$ may be different from $\sigma^2(\beta_0^T \mathbf{x})$ and $E(\mathbf{x}|\beta_0^T \mathbf{x})$. When $\tilde{\sigma}^2(\beta_0^T \mathbf{x}) = \sigma^2(\beta_0^T \mathbf{x})$ and $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x}) = E(\mathbf{x}|\beta_0^T \mathbf{x})$, the above estimating equation has the form

$$E[\{\varepsilon^2 - \sigma^2(\beta_0^T \mathbf{x})\} \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\}] = \mathbf{0}. \quad (2.3)$$

It is not difficult to verify that equation (2.3) will produce consistent estimate of β_0 . In other words, β_0 is a solution to $E[\{\varepsilon^2 - \sigma^2(\beta^T \mathbf{x})\} \{\mathbf{x} - E(\mathbf{x}|\beta^T \mathbf{x})\}] = \mathbf{0}$. An important virtue of the estimating equation (2.2) is that, as long as one of the functions $\tilde{\sigma}^2(\beta_0^T \mathbf{x})$ and $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$ is correctly specified, equation (2.2) will always yield consistent estimate of β_0 . In particular, without knowing the exact form of $\sigma^2(\cdot)$, let us assume that we specify it as $e\sigma^2(\cdot)$, which may be different from $\sigma^2(\cdot)$, and impose the working estimating equation as follows:

$$E[\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x})\} \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\}] = \mathbf{0}. \quad (2.4)$$

By the law of double expectation, equation (2.4) is equivalent to

$$E(E[\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x})\} \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\} | \mathbf{x}]) = \mathbf{0}.$$

To verify the above estimating equation yields consistent estimate of the true β_0 at the population level, we recall that $E(\varepsilon^2 | \mathbf{x}) = \sigma^2(\beta_0^T \mathbf{x})$ under the model assumption (1.1). Then the left hand side of (2.4) can be further reduced as

$$\begin{aligned} & E[E\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x}) | \mathbf{x}\} \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\}] \\ &= E[E\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x}) | \beta_0^T \mathbf{x}\} \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\}] \\ &= E[\{\varepsilon^2 - \tilde{\sigma}^2(\beta_0^T \mathbf{x})\} E\{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x}) | \beta_0^T \mathbf{x}\}], \end{aligned}$$

where the last term obviously equals to $\mathbf{0}$, indicating that equation (2.4) is able to produce consistent estimator of β_0 . The derivations above provide the motivation of (2.2), as it will

continue to yield consistent estimate of β_0 even if $\sigma^2(\cdot)$ is misspecified. As a special case, let $\tilde{\sigma}^2(\beta_0^T \mathbf{x}) = 0$ in (2.4) and we get

$$E[\{\varepsilon^2 \{\mathbf{x} - E(\mathbf{x}|\beta_0^T \mathbf{x})\}\}] = 0, \quad (2.5)$$

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), (i.e., $E(\mathbf{x}|\beta_0^T \mathbf{x})$ is a linear function of $\mathbf{x}$), then β_0 must be proportional to $\{\text{var}(\mathbf{x})\}^{-1} \text{cov}(\mathbf{x}, \varepsilon^2)$. This property simplifies solving (2.5) dramatically 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 will study this issue in the next section. Aside from this, we have seen that the estimating equation (2.5) is a specific member of the general class of estimating equations (2.2).

On the other hand, correct specification of $E(\mathbf{x}|\beta_0^T \mathbf{x})$ may be challenging in some situations. Even if all components in $E(\mathbf{x}|\beta_0^T \mathbf{x})$ can be correctly specified, calculating $E(\mathbf{x}|\beta_0^T \mathbf{x})$ is rather intensive when $\mathbf{x}$ is high dimensional. In order to simplify the calculation, suppose that we estimate $E(\mathbf{x}|\beta_0^T \mathbf{x})$ by $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$. Then equation (2.2) becomes

$$E[\{\varepsilon^2 - \sigma^2(\beta_0^T \mathbf{x})\} \{\mathbf{x} - \tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})\}] = 0, \quad (2.6)$$

which is true by recognizing the fact that $E(\varepsilon^2|\mathbf{x}) = \sigma^2(\beta_0^T \mathbf{x})$ and conditioning on $\mathbf{x}$. In another word, equation (2.2) will continue to yield consistent estimate of β_0 even if $E(\mathbf{x}|\beta_0^T \mathbf{x})$ is misspecified. For example, we can simply set $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x}) = 0$ and equation (2.6) becomes

$$E[\{\varepsilon^2 - \sigma^2(\beta_0^T \mathbf{x})\} \mathbf{x}] = 0. \quad (2.7)$$

We remark here that (2.5) and (2.7) are equivalent at the population level by noting that $E\{\varepsilon^2 E(\mathbf{x}|\beta_0^T \mathbf{x})\} = E\{E(\varepsilon^2|\beta_0^T \mathbf{x})\mathbf{x}\} = E\{\sigma^2(\beta_0^T \mathbf{x})\mathbf{x}\}$ under model (1.1).

We have shown that estimating equation (2.2) has a desirable robustness property: As long as either $\tilde{\sigma}^2(\cdot)$ or $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$ is correctly specified, the estimator based on the sample version of (2.2) will be consistent. Both parametric and nonparametric methods could be used to model $\tilde{\sigma}^2(\cdot)$ or $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$. For example, $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$ is typically assumed to be a linear function of $\beta_0^T \mathbf{x}$ in the dimension reduction literature (Li, 1991). We propose to estimate $\tilde{\sigma}^2(\cdot)$ and $\tilde{E}(\mathbf{x}|\beta_0^T \mathbf{x})$ 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 the independent and identically distributed samples $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n\}$, we now discuss sample versions of the proposed estimating equations and their asymptotic properties.

Ideally, one may estimate β_0 by solving for $\hat{\beta}$ that satisfies

$$n^{-1/2} \sum_{i=1}^n \{ \varepsilon_i^2 - \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i) \} \{ \mathbf{x}_i - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \} = \mathbf{0}, \quad (2.8)$$

which is the sample counterpart of (2.3). In (2.8), the quantities $\hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i)$ and $\hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i)$ are the consistent estimators of $\sigma^2(\beta^T \mathbf{x}_i)$ and $E(\mathbf{x}_i | \beta^T \mathbf{x}_i)$, 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 $\hat{\varepsilon}_i$. To guarantee the consistency of $\hat{\beta}$, we need consistent estimation of ε_i , indicating that we have to estimate $E(Y | \mathbf{x})$ consistently. Under model assumption (1.2) that $E(Y | \mathbf{x}) = \ell(\alpha_0^T \mathbf{x})$, we estimate α_0 by solving for $\hat{\alpha}$ that satisfies

$$n^{-1/2} \sum_{i=1}^n \{ Y_i - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) \} \{ \mathbf{x}_i - \hat{E}(\mathbf{x}_i | \hat{\alpha}^T \mathbf{x}_i) \} = \mathbf{0}, \quad (2.9)$$

which is parallel to (2.8) with

$$\hat{E}(\mathbf{x}_i | \hat{\alpha}^T \mathbf{x}_i) = \frac{\sum_{j=1}^n K_h(\hat{\alpha}^T \mathbf{x}_j - \hat{\alpha}^T \mathbf{x}_i) \mathbf{x}_j}{\sum_{j=1}^n K_h(\hat{\alpha}^T \mathbf{x}_j - \hat{\alpha}^T \mathbf{x}_i)}, \quad \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) = \frac{\sum_{j=1}^n K_{h_1}(\hat{\alpha}^T \mathbf{x}_j - \hat{\alpha}^T \mathbf{x}_i) Y_j}{\sum_{j=1}^n K_{h_1}(\hat{\alpha}^T \mathbf{x}_j - \hat{\alpha}^T \mathbf{x}_i)},$$

In the above denote $K_h(\cdot) = K(\cdot/h)/h$ as the kernel function, h and h_1 as the bandwidths.

Next we calculate the residual $\hat{\varepsilon}_i = Y_i - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i)$, and get our final estimate of β_0 by solving

$$n^{-1/2} \sum_{i=1}^n \{ \hat{\varepsilon}_i^2 - \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i) \} \{ \mathbf{x}_i - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \} = \mathbf{0}. \quad (2.10)$$

In (2.10) we estimate $E(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i)$ and $\sigma^2(\hat{\beta}^T \mathbf{x}_i)$ as follows,

$$\hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) = \frac{\sum_{j=1}^n K_h(\hat{\beta}^T \mathbf{x}_j - \hat{\beta}^T \mathbf{x}_i) \mathbf{x}_j}{\sum_{j=1}^n K_h(\hat{\beta}^T \mathbf{x}_j - \hat{\beta}^T \mathbf{x}_i)}, \quad \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i) = \frac{\sum_{j=1}^n K_{h_2}(\hat{\beta}^T \mathbf{x}_j - \hat{\beta}^T \mathbf{x}_i) \hat{\varepsilon}_j^2}{\sum_{j=1}^n K_{h_1}(\hat{\beta}^T \mathbf{x}_j - \hat{\beta}^T \mathbf{x}_i)}.$$

To summarize, we implement the following algorithm to estimate β_0

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

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

In case the matrix $J'(\alpha^{(k)})$ is singular or nearly so, we adopt a ridge regression approach as follows: use (2.11) with $J'(\alpha^{(k)})$ replaced by $J'_r(\alpha^{(k)}) = J'(\alpha^{(k)}) + \lambda_n \mathbf{I}_{p \times p}$ 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 $\alpha^{(k+1)}$ fails to change. Denote the resultant estimate by $\hat{\alpha}$.
2. Obtain $\hat{\ell}(\hat{\alpha}^T \mathbf{x}_i)$ by using kernel regression of Y_i onto $(\hat{\alpha}^T \mathbf{x}_i)$, for $i = 1, \dots, n$.
3. Solve (2.10) to obtain $\hat{\beta}$, where $\varepsilon_i = Y_i - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i)$.
 - a. Start with an initial value $\beta^{(0)}$. In our implementation, we choose $\beta^{(0)} = \left\{ \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \right\}^{-1} \left\{ \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) \varepsilon_i^2 \right\}$.
 - b. Write $I(\beta) = n^{-1/2} \sum_{i=1}^n \{ \varepsilon_i^2 - \hat{\sigma}^2(\beta^T \mathbf{x}_i) \} \{ \mathbf{x}_i - \hat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \}$ and $I'(\beta) = -n^{-1/2} \sum_{i=1}^n \{ \hat{\sigma}^2(\beta^T \mathbf{x}_i) \}' \{ \mathbf{x}_i - \hat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \} \mathbf{x}_i^T$, where the derivative $\{ \hat{\sigma}^2(\beta^T \mathbf{x}_i) \}' = \{ \sigma^2(\beta^T \mathbf{x}_i) \}' / (\beta^T \mathbf{x}_i)$ is taken from the kernel estimator $\sigma^2(\beta^T \mathbf{x}_i)$. Update $\beta^{(k)}$ with

$$\beta^{(k+1)} = \beta^{(k)} - \left\{ I'(\beta^{(k)}) + \lambda_n \mathbf{I}_{p \times p} \right\}^{-1} \{ I(\beta^{(k)}) \}. \quad (2.12)$$
 - c. Iterate (2.12) until $\beta^{(k+1)}$ fails to change. Denote the resultant estimate by $\hat{\beta}$.
4. Obtain $\hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i)$ by using kernel regression of ε_i^2 onto $(\hat{\beta}^T \mathbf{x}_i)$, for $i = 1, \dots, n$.

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.

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

$$\mathbf{Q} = E \left[\mathbf{A}(\mathbf{x}) \left\{ \sigma_0^2(\beta_0^T \mathbf{x}) \right\}' \mathbf{x}^T \right] \text{ and } \mathbf{V} = E \left\{ \text{var}(\varepsilon^2 | \mathbf{x}) \mathbf{A}(\mathbf{x}) \mathbf{A}^T(\mathbf{x}) \right\},$$

where $\mathbf{A}(\mathbf{x}) = \mathbf{x} - E(\mathbf{x} | \beta_0^T \mathbf{x})$. Then $n^{1/2} \mathbf{V}^{-1/2} \mathbf{Q} (\hat{\beta} - \beta_0)$ converges in distribution to the standard multivariate normal distribution as $n \rightarrow \infty$.

Theorem 1 provides the asymptotic normality of $\hat{\beta}$ obtained from step 3 in our proposed algorithm, which uses the residual ε_i and relies on equation (2.10). As detailed in the proof in the Appendix, we can obtain exactly the same asymptotic distribution of $\hat{\beta}$ based on equation (2.8) with known error ε_i . This means our procedure performs as well as the oracle procedure in terms of estimating β_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 $\hat{\beta}$, we estimate $\sigma^2(\cdot)$ via kernel regression. The next theorem states the consistency of $\hat{\sigma}^2(\cdot)$ obtained from step 4 in our proposed algorithm.

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

$$\text{bias} = h_2^2 \mu_2 \left[\left\{ \sigma^2(\beta_0^T \mathbf{x}) \right\}'' / 2 + \left\{ \sigma^2(\beta_0^T \mathbf{x}) \right\}' \left\{ f(\beta_0^T \mathbf{x}) \right\}' / f(\beta_0^T \mathbf{x}) \right],$$

where $\mu_2 = \int_{-1}^1 u^2 K(u) du$, $\{\sigma^2(\cdot)\}'$, $\{f(\cdot)\}'$ and $\{\sigma^2(\cdot)\}''$ denote the first and second order derivatives, respectively. Then $(nh_2)^{1/2} \{\hat{\sigma}^2(\hat{\beta}^T \mathbf{x}) - \sigma^2(\beta_0^T \mathbf{x}) - \text{bias}\}$ converges in distribution to normal distribution with mean zero and variance $\text{var}(\varepsilon^2 | \beta_0^T \mathbf{x}) / f(\beta_0^T \mathbf{x})$.

Because $\hat{\beta}$ 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 $\hat{\beta}^T \mathbf{x}$ as efficiently as if $\beta_0^T \mathbf{x}$ 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 β_0 from the estimating equation (2.10) and establish the adaptive property for $\hat{\beta}$ and $\hat{\sigma}^2(\hat{\beta}^T \mathbf{x})$ 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 derive an estimate for β_0 using the sample version of (2.5) or (2.7). However, it is necessary to undersmooth $\hat{E}(\mathbf{x} | \beta_0^T \mathbf{x})$ in (2.5) or $\hat{\sigma}^2(\hat{\beta}^T \mathbf{x})$ in (2.7) in order for the resulting estimate to achieve the adaptive properties. We skipped the details from the present context.

3. Numerical Studies

3.1 Simulations

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

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

Solving equation (3.1) will yield the classical nonlinear least squares estimation proposed by Härdle et al. (1993), and it serves as a benchmark for our comparison.

Estimating equations for the sample level of (2.7) are as follows:

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

We recall that the quantity $\tilde{E}(\mathbf{x} | \beta_0^T \mathbf{x})$ is misspecified to be $\mathbf{0}$ in (2.7). We have seen in Section 2.1 that the estimator of β_0 obtained from (3.2) remains to be consistent if $\hat{\varepsilon}^2$ and $\hat{\sigma}^2(\cdot)$ 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 given below:

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

In equations (3.1), (3.2) and (3.3), $\varepsilon_i = Y_i - \ell(\alpha^T \mathbf{x}_i)$. As defined in Section 2.2, we estimate $\ell(\alpha^T \mathbf{x})$, $\sigma^2(\beta^T \mathbf{x})$, $E(\mathbf{x} | \alpha^T \mathbf{x})$ and $E(\mathbf{x} | \beta^T \mathbf{x})$ through the corresponding kernel estimates. Both $\sigma^2(\beta^T \mathbf{x})$ and $E(\mathbf{x} | \beta^T \mathbf{x})$ 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 $\hat{\beta}$ holds true for a wide range of bandwidth.

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

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

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

Conditioning on $\mathbf{x}$, Y is generated from normal distribution with two different mean functions

- a. $\ell(\alpha_0^T \mathbf{x}) = 5(\alpha_0^T \mathbf{x})$,
- b. $\ell(\alpha_0^T \mathbf{x}) = 2\exp(\alpha_0^T \mathbf{x})$,

where $\alpha_0 = (0.8, 0.4, -0.4, 0.2, 0, 0, 0, 0)^T$ in (a) and (b) and the variance function $\sigma^2(\beta_0^T \mathbf{x}) = 0.25(\beta_0^T \mathbf{x} + 2)^2$ with $\beta_0 = (-0.45, 0, \dots, 0, 0.9)^T$.

Performances of estimating β_0 and β_0 . We first evaluate the estimation accuracy of β_0 and β_0 . The simulations were repeated 1000 times with sample size $n = 600$. The bias (“bias”) and the standard deviation (“std”) of the estimates of typical elements of β_0 and β_0 are reported in Tables 3.1 and 3.2 for Cases 1 and 2 respectively.

The estimating equations (3.1), (3.2) and (3.3) have similar performances for estimating $\hat{\alpha}$ in terms of both the bias and the standard deviation. The estimating equations (3.3) perform the best for estimating β_0 , and (3.2) perform only slightly worse, confirming the robustness property of the equation (2.2). The estimating equations (3.1) have the largest biases and standard deviations, which may be explained by the fact that estimation of $\{\ell(\cdot)\}'$ and $\{\sigma^2(\cdot)\}'$ are involved. From Tables 3.1 and 3.2, all three procedures are close to their corresponding oracle estimates, which adopt the true errors instead of the residuals. This confirms the our proposed estimators have the adaptive property.

Performances of estimating $\ell(\cdot)$ and $\sigma^2(\cdot)$ We next assess the performances in estimating $\ell(\cdot)$ and $\sigma^2(\cdot)$. We define the average squared errors criterion as follows,

$$\begin{aligned} ASE(\hat{\alpha}) &= n^{-1} \sum_{i=1}^n \left\{ \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) - \ell(\alpha_0^T \mathbf{x}_i) \right\}^2, \text{ and} \\ ASE(\hat{\beta}) &= n^{-1} \sum_{i=1}^n \left\{ \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i) - \sigma^2(\beta_0^T \mathbf{x}_i) \right\}^2. \end{aligned}$$

The results are reported in Table 3.3. 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 Zealand's South Island (Camden, 1989). The response variable Y is the shell mass, three quantitative predictors are 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), which are very similar to the results obtained from (3.1) and (3.2). By assuming (1.2) for the mean function, we estimate α_0 by the first equation of (3.3), and obtain that $\alpha = (0.3615, 0.4243, 0.8302)^T$. Based on $\{(\alpha^T \mathbf{x}_i, Y_i), i = 1, \dots, n\}$, we estimate $\ell(\alpha_0^T \mathbf{x})$ by kernel regression. The estimated regression function and its point-wise prediction interval are plotted in Figure 3.1(A). The prediction interval at the 95% level is calculated by assuming tentatively the variance function is a constant. We can clearly see that the empirical converge probability is very poor, particularly when $\alpha^T \mathbf{x}$ is large. In effect only 63 among 201 points lie within this prediction region. It follows that homoscedasticity is not a reasonable assumption for this data set.

Next we solve the second equation of the estimating equation (3.3) and obtain that $\hat{\beta} = (-0.3792, 0.3724, 0.8471)^T$. The variance function is then estimated by kernel regression based on data points $\{(\hat{\beta}^T \mathbf{x}_i, \hat{\varepsilon}_i^2), i = 1, \dots, n\}$. The estimated variance function is plotted in Figure 3.1(B), which does not seem to be constant. Taking into account the heteroscedasticity, we reported the 95% prediction interval of $b\ell(\hat{\beta}^T \mathbf{x})$ in Figure 3.1(C). We can see that, around 94% of the sample (189 points) was covered by this prediction interval, which is very close to 95%.

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 α and α and the link function $\ell(\cdot)$. We then bootstrap the original data 1000 times. Three quantities can be obtained from the bootstrap sample: α^* denotes the estimator of α_0 based on the bootstrap sample $(\mathbf{x}_i^*, Y_i^*)$; $\hat{\beta}^*$ denotes the estimator of β_0 based on $\{\mathbf{x}_i^*, Y_i^* - \hat{\ell}^*(\mathbf{x}_i^{*T} \hat{\alpha}^*)\}$ and $\hat{\beta}_0^*$ denotes the estimator of β_0 based on $\{\mathbf{x}_i^*, Y_i^* - \hat{\ell}(\mathbf{x}_i^{*T} \hat{\alpha})\}$. We remark here that $\hat{\beta}_0^*$ differs from $\hat{\beta}^*$ in that the former used the “true error” term because it adopted α and $\ell(\cdot)$ which were estimated from the original data, while the latter used the residuals calculated from the bootstrap sample. The third plot in Figure 3.1(D) present the boxplot of the absolute values of correlation coefficients $\text{corr}(\mathbf{x}^T \hat{\beta}^*, \mathbf{x}^T \hat{\beta})$. It implies that β_0^* behaves very similarly to $\hat{\beta}_0^*$ because the correlation coefficients are all very large. The first and second plots in Figure 3.1(D) present the respective boxplots of the absolute value of the correlation coefficients $\text{corr}(\mathbf{x}^T \alpha^*, \mathbf{x}^T \alpha)$ and $\text{corr}(\mathbf{x}^T \hat{\beta}^*, \mathbf{x}^T \hat{\beta})$. It can be seen that α^* performs much more stable than $\hat{\beta}^*$ and $\hat{\beta}_0^*$, indicating that estimating the conditional variance function is more difficult than estimating the conditional mean function.

4. Discussions

Estimation of the conditional heteroscedasticity remains an important and open problem in the literature when the dimension of predictors is very large (Antoniadis et al., 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 $\varepsilon = \sigma(\beta_0^T \mathbf{x})\varepsilon$. Thus, one can estimate β_0 by considering $|\varepsilon|^\alpha = \sigma^\alpha(\beta_0^T \mathbf{x})E|\varepsilon|^\alpha + \xi$ for any $\alpha > 0$, where $\xi = \sigma^\alpha(\beta_0^T \mathbf{x})(|\varepsilon|^\alpha - E|\varepsilon|^\alpha)$. 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 possibility for α rather than $\alpha = 2$, however, this is beyond the scope of this paper.

Acknowledgments

Zhu's research was supported by National Natural Science Foundation of China (NNSFC) 11071077 and the National Institute on Drug Abuse (NIDA) grant R21-DA024260. 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, and NIDA grant P50-DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the National Institutes of Health.

References

- Anderson TG, Lund J. Estimating continuous time stochastic volatility models of the short term interest rate. *J. Econometrics*. 1997; 77:343–377.
- Antoniadis A, Grégoire G, Mckeague I. Bayesian estimation in single-index models. *Stat. Sinica*. 2004; 14:1147–1164.
- Box GEP, Cox DR. An analysis of transformations (with discussion). *J. R. Statist. Soc. B*. 1964; 26:211–252.
- Cai T, Levine M, Wang L. Variance function estimation in multivariate nonparametric regression with fixed design. *J. Mult. Anal*. 2009; 100:126–136.
- Camden, M. The Data Bundle. WeHington, New Zealand: New Zealand Statistical Association; 1989.
- Carroll, RJ.; Ruppert, D. Transformations and Weighting in Regression. London: Chapman & Hall; 1988.
- Fan J, Yao Q. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*. 1998; 85:645–660.
- Hall P, Carroll RJ. Variance function estimating in regression: the effect of estimating the mean. *J. R. Statist. Soc. B*. 1989; 51:3–14.
- Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Ann. Stat.* 1993; 21:157–178.
- Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*. 1993; 58:71–120.
- Li B, Dong Y. Dimension reduction for non-elliptically distributed predictors. *Ann. Stat.* 2009; 37:1272–1298.
- Li K-C. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* 1991; 86:316–327.
- Li LX, Zhu LP, Zhu LX. Inference on primary parameter of interest with aid of dimension reduction estimation. *J. R. Statist. Soc. B*. 2011; 73:59–80.
- Mercurio D, Spokoiny V. Statistical inference for time-inhomogeneous volatility models. *Ann. Statist.* 2004; 32:577–602.
- Müller H-G, Stadtmüller U. Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* 1987; 15:610–625.
- Pollard, D. Convergence of Stochastic Processes. New York: Springer; 1984.
- Ruppert D, Wand MP. Multivariate locally weighted least squares regression. *Ann. Stat.* 1994; 22:1346–1370.

- Ruppert D, Wand M, Holst U, Hössjer O. Local polynomial variance function estimation. *Technometrics*. 1997; 39:262–273.
- Serfling, RJ. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons Inc.; 1980.
- Song Q, Yang L. Spline confidence bands for variance function. *Journal of Nonparametric Statistics*. 2009; 21:589–609.
- Wang L, Brown LD, Cai T, Levine M. Effect of mean on variance function estimation on nonparametric regression. *Ann. Stat.* 2008; 36:646–664.
- Xia Y, Tong H, Li WK. Single-index volatility models and estimation. *Stat. Sinica*. 2002; 12:785–799.
- Yao Q, Tong H. Quantifying the influence of initial values on nonlinear prediction. *J. R. Statist. Soc. B*. 1994; 56:701–725.
- Yin J, Geng Z, Li R, Wang H. Nonparametric covariance model. *Stat. Sinica*. 2010; 20:469–479.
- Zhu LX, Fang KT. Asymptotics for kernel estimate of sliced inverse regression. *Ann. Stat.* 1996; 24:1053–1068.

Appendix: Technical Conditions and Proofs

Appendix A: Some technical conditions

- (C1) The density functions of $(\alpha_0^T \mathbf{x})$ and $(\beta_0^T \mathbf{x})$, which are denoted by $f_{\alpha_0}(\alpha_0^T \mathbf{x})$ and $f_{\beta_0}(\beta_0^T \mathbf{x})$, are continuous and bounded from zero and above for all $\mathbf{x} \in \mathbb{R}^p$, and its second derivative is locally Lipschitz.
- (C2) The functions $\ell(\alpha_0^T \mathbf{x})$, $\sigma^2(\alpha_0^T \mathbf{x})$, $\ell(\alpha_0^T \mathbf{x})f_{\alpha_0}(\alpha_0^T \mathbf{x})$, $\sigma^2(\beta_0^T \mathbf{x})f_{\beta_0}(\beta_0^T \mathbf{x})$, $E(\mathbf{x}|\alpha_0^T \mathbf{x})f_{\alpha_0}(\alpha_0^T \mathbf{x})$ and $E(\mathbf{x}|\beta_0^T \mathbf{x})f_{\beta_0}(\beta_0^T \mathbf{x})$ 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 the support $[-1, 1]$, and is Lipschitz continuous.
- (C4) The bandwidths h , h_1 and h_2 satisfy that $nh^4 \rightarrow \infty$ and $nh^8 \rightarrow 0$, $nh_i^4 \rightarrow \infty$ and $nh_i^8 \rightarrow 0$ for $i = 1$ and 2 .
- (C5) $E(Y^4) < \infty$ and $E(\mathbf{x}_i^4) < \infty$ for $i = 1, \dots, p$. In addition, the conditional variance of Y given $\mathbf{x}$ is bounded away from 0 and above.

Appendix B: Proofs of Theorems 1 and 2

Proof of Theorem 1. In our subsequent derivations, there are several places involving $\hat{\ell}(\alpha^T \mathbf{x})$, $\hat{\sigma}^2(\beta^T \mathbf{x})$ and $\hat{E}(\mathbf{x}|\beta^T \mathbf{x})$. Before we prove Theorem 1, we first quantify to what extent these functions can approximate their respective true values. We take $\hat{\ell}(\alpha^T \mathbf{x})$ as an example. We note that

$$\hat{\ell}(\hat{\alpha}^T \mathbf{x}) - \ell(\alpha_0^T \mathbf{x}) = \left\{ \hat{\ell}(\alpha^T \mathbf{x}) - \ell(\alpha_0^T \mathbf{x}) \right\} + \left\{ \hat{\ell}(\hat{\alpha}^T \mathbf{x}) - \hat{\ell}(\alpha^T \mathbf{x}) \right\}.$$

The first term in the right hand side can be dealt with using the standard nonparametric regression theory. Therefore, it remains to quantify the difference $\hat{\ell}(\hat{\alpha}^T \mathbf{x}) - \hat{\ell}(\alpha^T \mathbf{x})$. Let $C = \{\alpha: \|\alpha - \alpha_0\| \leq Cn^{-1/2}\}$. Recall that Li et al. (2011, Lemma 1) proved that, if $nh_1^4 \rightarrow \infty$,

$$\sup_{\mathbf{x} \in \mathbb{R}^p} \sup_c \left| \left\{ \widehat{\ell}(\tilde{\alpha}^T \mathbf{x}) - \widehat{\ell}(\alpha_0^T \mathbf{x}) \right\} - E \left\{ \widehat{\ell}(\tilde{\alpha}^T \mathbf{x}) - \widehat{\ell}(\alpha_0^T \mathbf{x}) \right\} \right| = O_p \{ \log n / (nh_1^2) \}.$$

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 for proving Lemma 3.3 of Zhu and Fang (1996), we can show without much difficulty that

$$E \left\{ \widehat{\ell}(\alpha_0^T \mathbf{x}) \right\} - \ell(\alpha_0^T \mathbf{x}) - \mu_2 h_1^2 \{ \ell''(\alpha_0^T \mathbf{x}) + \ell'(\alpha_0^T \mathbf{x}) f'(\alpha_0^T \mathbf{x}) / f(\alpha_0^T \mathbf{x}) \} = O(h^4),$$

where $\mu_2 = \int_{-1}^1 u^2 K(u) du$. A similar result also holds when α_0 is replaced by $\tilde{\alpha}$. By the mean-value theorem, it follows that

$$\sup_c |\ell''(\tilde{\alpha}^T \mathbf{x}) - \ell''(\alpha_0^T \mathbf{x})| h_1^2 \mu_2 = O_p \left(h_1^2 / n^{1/2} \right) = O(h_1^4),$$

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

$$\widehat{\ell}(\alpha_0^T \mathbf{x}) - \widehat{\ell}(\tilde{\alpha}^T \mathbf{x}) = \ell'(\alpha_0^T \mathbf{x}) \mathbf{x}^T (\alpha_0 - \tilde{\alpha}) + o_p(n^{-1/2}). \quad (\text{B.1})$$

Following similar arguments, we can obtain that

$$\begin{aligned} \hat{\sigma}_0^2(\beta_0^T \mathbf{x}) - \hat{\sigma}_0^2(\tilde{\beta}^T \mathbf{x}) &= \{ \sigma_0^2(\beta_0^T \mathbf{x}) \}' \mathbf{x}^T (\beta_0 - \tilde{\beta}) + o_p(n^{-1/2}), \text{ and} \\ \widehat{E}(\mathbf{x} | \beta_0^T \mathbf{x}) - \widehat{E}(\mathbf{x} | \tilde{\beta}^T \mathbf{x}) &= \{ E(\mathbf{x} | \beta_0^T \mathbf{x}) \}' \mathbf{x}^T (\beta_0 - \tilde{\beta}) + o_p(n^{-1/2}). \end{aligned} \quad (\text{B.2})$$

(B.1) and (B.2) will be used repetitively in our subsequent derivations. We discuss the consistency of $\hat{\beta}$ first. Denote

$$\begin{aligned} I(\beta) &= E \{ \varepsilon^2 - \sigma^2(\beta^T \mathbf{x}) \} \{ \mathbf{x} - E(\mathbf{x} | \beta^T \mathbf{x}) \}, \\ \widehat{I}(\beta) &= n^{-1} \sum_{i=1}^n E \{ \hat{\varepsilon}_i^2 - \hat{\sigma}^2(\beta^T \mathbf{x}_i) \} \{ \mathbf{x}_i - \widehat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \}. \end{aligned}$$

Let $U(\beta_0)$ be any open set that includes β_0 . To prove the consistency of $\hat{\beta}$, we assume that $\inf_{\beta \in \Theta \setminus U(\beta_0)} |I(\beta)| \geq \delta$ for some positive constant δ . It suffices to show that

$$Pr \left\{ \inf_{\beta \in \Theta \setminus U(\beta_0)} |\widehat{I}(\beta)| \leq \delta/2 \right\} \rightarrow 0. \quad (\text{B.3})$$

The condition $\inf_{\beta \in \Theta \setminus U(\beta_0)} |I(\beta)| \geq \delta$ implies that

$$Pr \left\{ \inf_{\beta \in \Theta \setminus U(\beta_0)} |\widehat{I}(\beta)| \leq \delta/2 \right\} \leq Pr \left\{ \inf_{\beta \in \Theta \setminus U(\beta_0)} |\widehat{I}(\beta) - I(\beta)| \geq \delta/2 \right\}.$$

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

$$Pr \left\{ \sup_{\beta \in \Theta} |\hat{I}(\beta) - I(\beta)| \geq \delta/2 \right\} \rightarrow 0.$$

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

$$I_{1,1} = n^{-1} \sum_{i=1}^n \{ \varepsilon_i^2 - \sigma^2(\beta^T \mathbf{x}_i) \} \{ \mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i) \} - I(\beta),$$

$$I_{1,2} = n^{-1} \sum_{i=1}^n \{ \hat{\varepsilon}_i^2 - \varepsilon_i^2 \} \{ \mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i) \}$$

$$I_{1,3} = n^{-1} \sum_{i=1}^n \{ \sigma^2(\beta^T \mathbf{x}_i) - \hat{\sigma}^2(\beta^T \mathbf{x}_i) \} \{ \mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i) \},$$

$$I_{1,4} = n^{-1} \sum_{i=1}^n \{ \varepsilon_i^2 - \sigma^2(\beta^T \mathbf{x}_i) \} \left\{ E(\mathbf{x}_i | \beta^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \right\},$$

$$I_{1,5} = n^{-1} \sum_{i=1}^n \{ \hat{\varepsilon}_i^2 - \varepsilon_i^2 \} \left\{ E(\mathbf{x}_i | \beta^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \right\},$$

$$I_{1,6} = n^{-1} \sum_{i=1}^n \{ \sigma^2(\beta^T \mathbf{x}_i) - \hat{\sigma}^2(\beta^T \mathbf{x}_i) \} \left\{ E(\mathbf{x}_i | \beta^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta^T \mathbf{x}_i) \right\}.$$

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 $\sup_{\beta \in \Theta} |I_{1,1}| = o_p(\log n / n^{1/2})$ almost surely (Pollard, 1984, Theorem 2.37).

$$\begin{aligned} I_{1,2} &= 2n^{-1} \sum_{i=1}^n \left[\left\{ \ell(\alpha_0^T \mathbf{x}_i) - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) \right\} \varepsilon_i \{ \mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i) \} \right] \\ &\quad + n^{-1} \sum_{i=1}^n \left[\left\{ \ell(\alpha_0^T \mathbf{x}_i) - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) \right\}^2 \{ \mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i) \} \right] = 2I_{1,2,1} + I_{1,2,2}. \end{aligned}$$

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

$$\begin{aligned}
|I_{1,2,1}| &\leq \sup_{\mathbf{x} \in \mathbb{R}^p} \left| \ell(\alpha_0^T \mathbf{x}) - \hat{\ell}(\alpha_0^T \mathbf{x}) \right| n^{-1} \sum_{i=1}^n |\varepsilon_i \{\mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i)\}| + n^{-1} \sum_{i=1}^n \left| \left\{ \hat{\ell}(\alpha_0^T \mathbf{x}_i) - \hat{\ell}(\hat{\alpha}^T \mathbf{x}_i) \right\} \right| |\varepsilon_i \{\mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i)\}| \\
&\leq \sup_{\mathbf{x} \in \mathbb{R}^p} \left| \ell(\alpha_0^T \mathbf{x}) - \hat{\ell}(\alpha_0^T \mathbf{x}) \right| n^{-1} \sum_{i=1}^n |\varepsilon_i \{\mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i)\}| + 2n^{-1} \sum_{i=1}^n |\varepsilon_i \ell'(\alpha_0^T \mathbf{x}_i) \{\mathbf{x}_i - E(\mathbf{x}_i | \beta^T \mathbf{x}_i)\} \mathbf{x}_i^T| |\alpha_0 - \hat{\alpha}|.
\end{aligned}$$

Therefore, $\sup_{\beta \in \Theta} |I_{1,2,1}| = o_p(1)$ follows immediately from the consistency of $\hat{\alpha}$ and the uniform consistency of $\hat{\ell}(\alpha_0^T \mathbf{x})$. Similarly, $\sup_{\beta \in \Theta} |I_{1,2,2}| = o_p(1)$ can be proved. These two results imply that $\sup_{\beta \in \Theta} |I_{1,2}| = o_p(1)$. Using the U -statistic theory (Serfling, 1980), we can obtain that $\sup_{\beta \in \Theta} |I_{1,i}| = o_p(1)$, $i = 3, 4$. Following similar arguments to deal with $I_{1,2}$, we can have $\sup_{\beta \in \Theta} |I_{1,5}| = o_p(1)$. $\sup_{\beta \in \Theta} |I_{1,6}| = o_p(1)$ follows directly from the Cauchy-Schwartz inequality and the standard results of the uniform convergence of $\hat{E}(\mathbf{x} | \beta^T \mathbf{x})$ and $\sigma^2(\beta^T \mathbf{x})$ in nonparametric regression. By combining these results together, it follows that

$$Pr \left\{ \sup_{\beta \in \Theta} |\hat{I}(\beta) - I(\beta)| \geq \delta/2 \right\} \rightarrow 0$$

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

Next we examine the root- n consistency of $\hat{\beta}$. We expand the estimating equation as follows.

$$0 = n^{-1/2} \sum_{i=1}^n \{\mathbf{x}_i - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i)\} \{\hat{\varepsilon}_i^2 - \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}_i)\} =: \sum_{i=1}^{12} I_{2,i}. \text{ The terms } I_{2,i} \text{'s will be explicitly defined in the sequel.}$$

By the classical central limit theorem,

$$I_{2,1} =: n^{-1/2} \sum_{i=1}^n \{\mathbf{x}_i - E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i)\} \{\varepsilon_i^2 - \sigma^2(\beta_0^T \mathbf{x}_i)\} = O_p(1).$$

Invoking the root- n consistency of $\hat{\alpha}$, we can show without much difficulty that

$$I_{2,2} =: n^{-1/2} \sum_{i=1}^n \{\mathbf{x}_i - E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i)\} \{\hat{\varepsilon}_i^2 - \varepsilon_i^2\} = o_p(1).$$

Invoking (B.2), it follows that

$$I_{2,3} = \left[E\{\sigma^2(\beta_0^T \mathbf{x})\}' \{\mathbf{x} - E(\mathbf{x} | \beta_0^T \mathbf{x})\} \mathbf{x}^T + o_p(1) \right] n^{1/2} (\beta_0 - \hat{\beta}).$$

Because $E\{\mathbf{x}_i - E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) | \beta_0^T \mathbf{x}_i\} = 0$, the U -statistic theory implies

$$I_{2,4} =: n^{-1/2} \sum_{i=1}^n \{\mathbf{x}_i - E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i)\} \{\sigma^2(\beta_0^T \mathbf{x}) - \hat{\sigma}^2(\beta_0^T \mathbf{x})\} = o_p(1),$$

$$I_{2,5} =: n^{-1/2} \sum_{i=1}^n \left\{ E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) \right\} \{ \varepsilon_i^2 - \sigma^2(\beta_0^T \mathbf{x}) \} = o_p(1).$$

By the uniform convergence of $\hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i)$ and (B.1), it is easy to see

$$I_{2,6} =: n^{-1/2} \sum_{i=1}^n \left\{ E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) \right\} \{ \hat{\varepsilon}_i^2 - \varepsilon_i^2 \} = o_p(1),$$

$$I_{2,7} =: n^{-1/2} \sum_{i=1}^n \left\{ E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) \right\} \left\{ \hat{\sigma}^2(\beta_0^T \mathbf{x}) - \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}) \right\} = o_p(1).$$

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

$$I_{2,8} =: n^{-1/2} \sum_{i=1}^n \left\{ E(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) \right\} \{ \sigma^2(\beta_0^T \mathbf{x}) - \hat{\sigma}^2(\beta_0^T \mathbf{x}) \} = 0.$$

By using the fact that $E\{\varepsilon_i^2 - \sigma^2(\beta_0^T \mathbf{x}) | \mathbf{x}\} = 0$, we can show that

$$I_{2,9} =: n^{-1/2} \sum_{i=1}^n \left\{ \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \right\} \{ \varepsilon_i^2 - \sigma^2(\beta_0^T \mathbf{x}) \},$$

$$I_{2,10} =: n^{-1/2} \sum_{i=1}^n \left\{ \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \right\} \{ \hat{\varepsilon}_i^2 - \varepsilon_i^2 \} = o_p(1) n^{1/2} (\beta_0 - \hat{\beta}).$$

Invoking (B.2) again, we have

$$I_{2,11} =: n^{-1/2} \sum_{i=1}^n \left\{ \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \right\} \left\{ \hat{\sigma}^2(\beta_0^T \mathbf{x}) - \hat{\sigma}^2(\hat{\beta}^T \mathbf{x}) \right\} = o_p(1) n^{1/2} (\beta_0 - \hat{\beta}),$$

$$I_{2,12} =: n^{-1/2} \sum_{i=1}^n \left\{ \hat{E}(\mathbf{x}_i | \beta_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \hat{\beta}^T \mathbf{x}_i) \right\} \{ \sigma^2(\beta_0^T \mathbf{x}) - \hat{\sigma}^2(\beta_0^T \mathbf{x}) \} = o_p(1) n^{1/2} (\beta_0 - \hat{\beta}).$$

By combining the results above, we obtain that $(\beta_0 - \hat{\beta}) = O_p(n^{-1/2})$.

The asymptotic normality follows immediately from the root- n consistency described above and the classical central limit theorem. That is,

$$E \left[\{ \mathbf{x} - E(\mathbf{x} | \beta_0^T \mathbf{x}) \} \{ \sigma_0^2(\beta_0^T \mathbf{x}) \}' \mathbf{x}^T \right] n^{1/2} (\hat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^n \{ \mathbf{x}_i - E(\mathbf{x} | \beta_0^T \mathbf{x}_i) \} \{ \varepsilon_i^2 - \sigma^2(\beta_0^T \mathbf{x}_i) \} + o_p(1),$$

which completes the proof of Theorem 1.

Proof of Theorem 2. It can be proved following from (B.2), the root- n consistency of $\hat{\alpha}$ and $\hat{\beta}$, 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 $\hat{\alpha}$, its convergence rate and the asymptotic normality. First we prove the consistency of $\hat{\alpha}$. We define

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

Let $U(\beta_0)$ be any open set that includes β_0 . To prove the consistency of $\hat{\alpha}$, we assume that $\inf_{\alpha \in \Theta \setminus U(\alpha_0)} |J(\alpha)| \geq \delta$ for some positive constant δ . It suffices to show that

$$Pr \left\{ \inf_{\alpha \in \Theta \setminus U(\alpha_0)} |\hat{J}(\alpha)| \leq \delta/2 \right\} \rightarrow 0. \quad (C.1)$$

The condition $\inf_{\alpha \in \Theta \setminus U(\alpha_0)} |J(\alpha)| \geq \delta$ implies that

$$Pr \left\{ \inf_{\alpha \in \Theta \setminus U(\alpha_0)} |\hat{J}(\alpha)| \leq \delta/2 \right\} \leq Pr \left\{ \inf_{\alpha \in \Theta \setminus U(\alpha_0)} |\hat{J}(\alpha) - J(\alpha)| \geq \delta/2 \right\}.$$

Therefore, it suffices to show that

$$Pr \left\{ \sup_{\alpha \in \Theta} |\hat{J}(\alpha) - J(\alpha)| \geq \delta/2 \right\} \rightarrow 0$$

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

Next we prove the root- n consistency of $\hat{\alpha}$. With probability close to 1, by Taylor expansion, for α between β_0 and $\hat{\alpha}$,

$$\begin{aligned} 0 &= \hat{J}(\hat{\alpha}) = \hat{J}(\alpha_0) + \hat{J}'(\tilde{\alpha})(\hat{\alpha} - \alpha_0) \\ &= \hat{J}(\alpha_0) + \left\{ \hat{J}'(\tilde{\alpha}) - \hat{J}'(\alpha_0) \right\} (\hat{\alpha} - \alpha_0) + \hat{J}'(\alpha_0)(\hat{\alpha} - \alpha_0). \end{aligned} \quad (C.2)$$

Recall the definition of $\hat{\mathcal{J}}(\alpha)$. We expand $\hat{\mathcal{J}}(\alpha_0) \stackrel{\text{def}}{=} \sum_{i=1}^4 \hat{\mathcal{J}}_i$, where

$$\begin{aligned}\hat{\mathcal{J}}_1 &= n^{-1} \sum_{i=1}^n \left\{ Y_i - \ell(\alpha_0^T \mathbf{x}_i) \right\} \left\{ \mathbf{x}_i - E(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) \right\}, \\ \hat{\mathcal{J}}_2 &= n^{-1} \sum_{i=1}^n \left\{ Y_i - \ell(\alpha_0^T \mathbf{x}_i) \right\} \left\{ E(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) \right\}, \\ \hat{\mathcal{J}}_3 &= n^{-1} \sum_{i=1}^n \left\{ \ell(\alpha_0^T \mathbf{x}_i) - \hat{\ell}(\alpha_0^T \mathbf{x}_i) \right\} \left\{ \mathbf{x}_i - E(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) \right\}, \\ \hat{\mathcal{J}}_4 &= n^{-1} \sum_{i=1}^n \left\{ \ell(\alpha_0^T \mathbf{x}_i) - \hat{\ell}(\alpha_0^T \mathbf{x}_i) \right\} \left\{ E(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) - \hat{E}(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) \right\}.\end{aligned}$$

Because $E\{Y - \ell(\alpha_0^T \mathbf{x})\} \{ \mathbf{x} - E(\mathbf{x} | \alpha_0^T \mathbf{x}) \} = 0$, $\hat{\mathcal{J}}_1$ is of the order $O_p(n^{-1/2})$ by the classical central limit theorem. Both $\hat{\mathcal{J}}_2$ and $\hat{\mathcal{J}}_3$ are of the order $o_p(n^{-1/2})$ according to the standard U -statistics theory (Serfling, 1980); by using Cauchy-Schwartz inequality, $\hat{\mathcal{J}}_4$ is less than

$$\left[\hat{E} \left\{ \left\{ \ell(\alpha_0^T \mathbf{x}) - \hat{\ell}(\alpha_0^T \mathbf{x}) \right\}^2 \right\} \right]^{1/2} \left[E \left\{ \left\{ E(\mathbf{x} | \alpha_0^T \mathbf{x}) - \hat{E}(\mathbf{x} | \alpha_0^T \mathbf{x}) \right\}^2 \right\} \right]^{1/2},$$

which is of order $O_p\{h_1^4 + \log n / (nh_1)\}$. If $nh_1^8 \rightarrow 0$ and $nh_1^2 / \log n \rightarrow \infty$, then the last term is also $o_p(n^{-1/2})$. Combining the above results, we obtained that $\hat{\mathcal{J}}(\alpha_0) = O_p(n^{-1/2})$.

By using the weak law of large numbers, we can see that $\hat{\mathcal{J}}'(\alpha_0)$ converges in probability to $J'(\alpha_0) = E[\ell'(\alpha_0^T \mathbf{x}) \{ \mathbf{x} - E(\mathbf{x} | \alpha_0^T \mathbf{x}) \} \mathbf{x}^T]$ for $nh_1^2 \rightarrow \infty$ and $nh_1^8 \rightarrow 0$. Similarly, $\hat{\mathcal{J}}'(\hat{\alpha}) - \hat{\mathcal{J}}'(\alpha_0) = O_p(1)$ in that $\hat{\alpha}$ is a consistent estimate of α_0 . Therefore, $\hat{\alpha}$ is a root- n consistent estimate of α_0 .

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

$$n^{1/2}(\hat{\alpha} - \alpha_0) = \{J'(\alpha_0)\}^{-1} n^{-1/2} \sum_{i=1}^n \varepsilon_i \left\{ \mathbf{x}_i - E(\mathbf{x}_i | \alpha_0^T \mathbf{x}_i) \right\} + o_p(1)$$

where $J'(\alpha_0) = E[\ell'(\alpha_0^T \mathbf{x}) \{ \mathbf{x} - E(\mathbf{x} | \alpha_0^T \mathbf{x}) \} \mathbf{x}^T]$ and $\varepsilon_i = Y_i - \ell(\alpha_0^T \mathbf{x}_i)$.

(A): Homoscedastic Single Index Model

(B): The Estimated Variance Function

(C): Heteroscedastic Single Index Model

(D): Boxplot of Correlation Coefficients

Figure 3.1.

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

Table 3.1

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

method	mean estimate				variance estimate				oracle estimation			
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_8$	$\hat{\beta}_8$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_8$	$\hat{\beta}_8$
(a): $\ell(\alpha_0^T \mathbf{x}) = 5(\alpha_0^T \mathbf{x})$												
(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): $\ell(\alpha_0^T \mathbf{x}) = 2 \exp(\alpha_0^T \mathbf{x})$												
(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		

Table 3.2

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

method	mean estimate				variance estimate				oracle estimation			
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_8$		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_8$	
(a): $\ell(\alpha_0^T \mathbf{x}) = 5(\alpha_0^T \mathbf{x})$												
(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): $\ell(\alpha_0^T \mathbf{x}) = 2 \exp(\alpha_0^T \mathbf{x})$												
(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 3.3

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

model	method	Case 1		Case 2	
		$ASE(\hat{\alpha})$	$ASE(\hat{\beta})$	$ASE(\hat{\alpha})$	$ASE(\hat{\beta})$
(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