Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2015 January 1; 77(1): 171–194. doi:10.1111/rssb.12066.

Variance Function Partially Linear Single-Index Models¹

HENG LIAN,

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore. hengliao@ntu.edu.sg

HUA LIANG¹, and

Department of Statistics, George Washington University, Washington, D.C. 20052, U.S.A.

RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA
carroll@stat.tamu.edu

Abstract

We consider heteroscedastic regression models where the mean function is a partially linear single index model and the variance function depends upon a generalized partially linear single index model. We do not insist that the variance function depend only upon the mean function, as happens in the classical generalized partially linear single index model. We develop efficient and practical estimation methods for the variance function and for the mean function. Asymptotic theory for the parametric and nonparametric parts of the model is developed. Simulations illustrate the results. An empirical example involving ozone levels is used to further illustrate the results, and is shown to be a case where the variance function does not depend upon the mean function.

Keywords

Asymptotic theory; Estimating equation; Identifiability; Kernel regression; Modeling ozone levels; Partially linear single index model; Semiparametric efficiency; Single-index model; Variance function estimation

1 Introduction

We consider heteroscedastic regression models where the mean function is a partially linear single index model and the variance function depends upon a generalized partially linear single index model. We do not insist that the variance function depend only upon the mean function, as happens in the classical generalized partially linear single index model. Our model is

$$Y = \mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{m}_\mu) + g\{v(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \mathbf{m}_v)\} \epsilon, \quad (1)$$

¹Corresponding Author: Hua Liang (hliang@gwu.edu).

$$\mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{m}_\mu) = m_\mu(\mathbf{X}^T \boldsymbol{\alpha}) + \mathbf{X}^T \boldsymbol{\beta}, \quad (2)$$

$$v(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \mathbf{m}_v) = m_v(\mathbf{X}^T \boldsymbol{\theta}) + \mathbf{X}^T \boldsymbol{\zeta}, \quad (3)$$

where $g(\cdot)$ is a known function, while $m_\mu(\cdot)$ and $m_v(\cdot)$ are two unknown smooth functions, ε is independent of $\mathbf{X}$, $E(\varepsilon) = 0$ and $E(\varepsilon^2) = 1$, the last condition for identifiability. Generally, either $g(v) = v$ or $g(v) = \exp(v)$. We need additional restrictions on the parameters to ensure identifiability, specifically $\|\boldsymbol{\alpha}\| = \|\boldsymbol{\theta}\| = 1$ and $\boldsymbol{\alpha}^T \boldsymbol{\beta} = \boldsymbol{\theta}^T \boldsymbol{\zeta} = 0$.

This model retains the flexibility of a nonparametric regression model but has dimension reduction ability to avoid fitting a multivariate nonparametric regression function. Methods for estimating the mean function $\mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)$ in (1)-(2) are already well-established if the potential heteroscedasticity is ignored, and the linear and index components are different. For example, Yu and Ruppert (2002) proposed a penalized spline estimation procedure. Xia and Härdle (2006) integrated the dimension reduction idea and minimum average variance estimation (MAVE, Xia *et al.*, 2002). The partially linear structure specified in (2) is an important, practical, special case of a multiple-index structure for the mean function models, e.g., Xia (2008), and general models studied in dimension reduction, e.g., Ma and Zhu (2013a). More recently, Wang *et al.* (2010) proposed a dimension-reduction based estimation procedure with additional slightly stronger assumptions, while Liang *et al.* (2010) proposed a profile least squares estimation procedure. While they allowed heteroscedasticity, they did not explicitly model it, instead they simply assumed that $E(Y|\mathbf{X}, \mathbf{Z}) = m_\mu(\mathbf{X}^T \boldsymbol{\alpha}) + \mathbf{Z}^T \boldsymbol{\beta}$, and their asymptotic theory assumed that $v(\mathbf{X}, \mathbf{Z}) \equiv 1$, although this improvement only changes their asymptotic covariance matrix. Moreover, all existing work mainly focuses on estimation of the mean parameters $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$.

Often, variance functions are thought to be nuisance parameters used only to improve the estimation of the mean, but this is a narrow view. There are many reasons why estimating variance can be important, and this has been understood for decades. Box and Hill (1974), in studies of kinetic rate parameters, used variance function estimation to improving understanding of model fit. Box and Meyer (1986) noted the crucial purpose of understanding variability as a function of covariates in discussing off-line quality control. Carroll and Ruppert (1988); Carroll (2003); Davidian and Carroll (1987); Davidian *et al.* (1988), all described the crucial role of understanding variability in calibration experiments for such contexts as assays. Cai and Wang (2008) stated that “In addition to being of interest in its own right, variance function estimates are needed, for example, to construct confidence intervals/bands for the mean function”. Western and Bloome (2009), in their study of social inequality, fitted a variance function model as the main purpose of their work, testing whether men recently released from a prison experience greater income insecurity, i.e., greater variance, in addition to the well documented decline in average earnings.

In the context of modeling heteroscedasticity in linear or nonlinear models (Bickel, 1978; Carroll, 1982; Carroll and Ruppert, 1982), there are two analysis strategies: a parametric

approach, in which the variance function is assumed to a parametric function of the covariates; and a nonparametric approach (Carroll and Härdle, 1989; Fuller and Rao, 1978; Hall and Carroll, 1989) with a fully nonparametric variance structure. This approach is hampered by the curse of dimensionality in practical applications. Ma *et al.* (2006) studied semiparametric efficiency in heteroscedastic partially linear models where $\mathbf{X}$ is scalar, a special case of our model. Van Keilegom and Wang (2010) studied a general class of location-dispersion regression models, including semiparametric quantile heteroscedastic regression. Their results are very general in terms of asymptotic conditions and theory, and encompass a wide variety of possible methods. However, the examples they use to illustrate their methods are special cases of our model (1)-(3). In addition, our methods have the practical advantage of being based numerically on a single algorithm, fitting a mean model $g\{m_*(\mathbf{X}^T\mathbf{a}) + \mathbf{X}^T\mathbf{b}\}$ with $\|\mathbf{a}\| = 1$ and $\mathbf{a}^T\mathbf{b} = 0$ and incorporating weights, see Section 2.7. In contrast, one of the variance function methods in Van Keilegom and Wang (2010) in our context involves a non-differentiable objective function. Ma and Zhu (2013a) applied the strategy developed in Ma *et al.* (2006) to heteroscedastic partially linear single index models and established doubly robust and efficient estimators of the mean parameters. They also developed a type of generalized least squares procedure for achieving further efficiency in estimating the mean function, but their variance models were very special cases of our model (3).

More recently, there is a recognition that variability itself can be a predictor of other outcomes. Thomas *et al.* (2012) showed in a context different from ours that individual variability in longitudinal measurements for an individual can be predictive of a health outcome. Teschendorff and Widschwendter (2012) argued that in cancer genomics, differential variability can be as important as differential means for predicting disease phenotypes. While the techniques in these two papers are different from ours, they indicate that understanding variability can be crucial as does the work of Western and Bloome (2009).

Our primary goal is to develop efficient and practical estimates the parameters θ and ζ , and then the variance function $g\{v(\mathbf{X}, \theta, \zeta, m_v)\}$. As a by-product, we can do re-weighting (generalized least squares) to improve the estimates of (α, β) . It is worth mentioning that the model that $\alpha = \theta, \beta = \zeta$ and $m_\mu(\cdot) = m_v(\cdot)$ is the special case that the variance is entirely a function of the mean, as is typical in GLIM's. The model (1)-(3) is also interesting conceptually because it suggests that if either $\alpha = \theta$ or $\beta = \zeta$, then different linear combinations of $\mathbf{X}$ are governing the distribution of Y . This can be important in classification, for example, because then there is extra information available when modeling Y . Also, we can use this to look for clusters of individuals who have excess variability: that excess variability can be of other interest.

The paper is organized as follows. Section 2 describes the estimation procedures and sketches the computational algorithms. Section 3 presents the main theoretical results and their implications. Section 4 presents the results of simulation studies and an analysis of ozone data, where it is shown that the variance function does not depend only upon the mean function. All technical assumptions and proofs of the main results are located in an Appendix.

2 Estimation Methods

2.1 Outline of the Estimation Methods

Our main goal, and our main innovation, is to develop efficient, practical, estimates of the variance function, along with their asymptotic theory. Along the way, we will of course develop efficient estimation of the mean function, although such estimation is much better understood than variance function estimation, as seen from the literature review in Section 1. Our development proceeds in four steps.

- Initial estimation of the mean function ignoring the heteroscedasticity.
- Using the absolute residuals from the initial estimate of the mean function, develop an initial estimate of the variance function.
- Update the mean function by using weights based on the initial variance function. This estimate is semiparametric efficient in the case that ε is normally distributed, and is shown in simulations to dominate the unweighted estimator.
- Update the variance function by using the absolute residuals from the weighted mean function. This method is shown in simulations to dominate the initial estimate of the variance function.

2.2 Preliminaries

The two constraints $\|\alpha\| = 1$ (with first nonzero component positive) and $\alpha^T \beta = 0$ in effect reduce the number of parameters $(\alpha^T, \beta^T)^T$ from $2p$ to $2p - 2$. Therefore, the popular “delete-one-component” method (Yu and Ruppert, 2002) can be used here. Without loss of generality, we assume the first component of α , α_1 is positive, and thus we can write $\alpha = ((1 - \|\alpha_{\setminus 1}\|^2)^{1/2}, \alpha_2, \dots, \alpha_p)^T$ where $\alpha_{\setminus 1} = (\alpha_2, \dots, \alpha_p)^T$ is α without the first component. Similarly we denote $\beta_{\setminus 1} = (\beta_2, \dots, \beta_p)^T$. Due to the second constraint, we have

$\beta_1 = -\alpha_1^{-1} \left(\sum_{j=2}^p \alpha_j \beta_j \right) = - \left(1 - \|\alpha_{\setminus 1}\|^2 \right)^{-1/2} \left(\sum_{j=2}^p \alpha_j \beta_j \right)$. Thus (α, β) is a function of $\omega := (\alpha_2, \dots, \alpha_p, \beta_2, \dots, \beta_p)^T$. The $2p \times (2p - 2)$ Jacobian matrix, $(\alpha^T, \beta^T)^T / \omega$, is

$$\mathbf{J}_\omega = \begin{pmatrix} -\frac{\alpha_{\setminus 1}^T}{(1 - \|\alpha_{\setminus 1}\|^2)^{1/2}} & 0_{1 \times (p-1)} \\ I_{(p-1) \times (p-1)} & 0_{(p-1) \times (p-1)} \\ -\frac{\beta_{\setminus 1}^T}{(1 - \|\alpha_{\setminus 1}\|^2)^{1/2}} - \left(\alpha_{\setminus 1}^T \beta_{\setminus 1} \right) \frac{\alpha_{\setminus 1}^T}{(1 - \|\alpha_{\setminus 1}\|^2)^{3/2}} & -\frac{\alpha_{\setminus 1}^T}{(1 - \|\alpha_{\setminus 1}\|^2)^{1/2}} \\ 0_{(p-1) \times (p-1)} & I_{(p-1) \times (p-1)} \end{pmatrix},$$

where we have indicated the dimensions of the zero matrices and identity matrices for clarity.

In the same way, we assume the first component of θ , θ_1 is positive, and thus we can introduce $\theta_{\setminus 1}$ and $\zeta_{\setminus 1}$. Write $\vartheta = (\theta_2, \dots, \theta_p, \zeta_2, \dots, \zeta_p)^T$, and define $\mathbf{J}_\vartheta$ similarly to $\mathbf{J}_\omega$.

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$ be independent samples of $(Y, \mathbf{X})$. In what follows, we set $R_i = |Y_i - \mu(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)|$, $\varepsilon_i = Y_i - \{m_\mu(\mathbf{X}_i^T \boldsymbol{\alpha}) + \mathbf{X}_i^T \boldsymbol{\beta}\}$, $d_i = I_{(\varepsilon_i > 0)} - I_{(\varepsilon_i \leq 0)} = \text{sign}(\varepsilon_i)$, $g_i = g\{m_v(\mathbf{X}_i^T \boldsymbol{\theta}) + \mathbf{X}_i^T \boldsymbol{\zeta}\}$, $\delta_i = R_i - g_i$, $U_i = \mathbf{X}_i^T \boldsymbol{\alpha}$, $T_i = \mathbf{X}_i^T \boldsymbol{\theta}$. Then $E(|\varepsilon_i| | \mathbf{X}_i) = E(R_i | \mathbf{X}_i) = g\{m_v(\mathbf{X}_i^T \boldsymbol{\theta}) + \mathbf{X}_i^T \boldsymbol{\zeta}\}$. Let $g^{(1)}$ denote the first derivative of g , define $g^{(1)2} = \{g^{(1)}\}^2$. Set $\Lambda_\mu = \{m_\mu^{(1)}(\mathbf{X}^T \boldsymbol{\alpha}) \mathbf{X}^T, \mathbf{X}^T\}^T$, $\Lambda_{i\mu} = \{m_\mu^{(1)}(\mathbf{X}_i^T \boldsymbol{\alpha}) \mathbf{X}_i^T, \mathbf{X}_i^T\}$, $\Lambda_v = \{m_v^{(1)}(\mathbf{X}^T \boldsymbol{\theta}) \mathbf{X}^T, \mathbf{X}^T\}^T$, $\Lambda_{iv} = \{m_v^{(1)}(\mathbf{X}_i^T \boldsymbol{\theta}) \mathbf{X}_i^T, \mathbf{X}_i^T\}$, $\tilde{\Lambda}_\mu = \Lambda_\mu - E(\Lambda_\mu / g^2 | \mathbf{X}^T \boldsymbol{\alpha}) / E(1/g^2 | \mathbf{X}^T \boldsymbol{\alpha})$, $\tilde{\Lambda}_{i\mu} = \Lambda_{i\mu} - E(\Lambda_{i\mu} / g_i^2 | \mathbf{X}_i^T \boldsymbol{\alpha}) / E(1/g_i^2 | \mathbf{X}_i^T \boldsymbol{\alpha})$, $\tilde{\Lambda}_v = \Lambda_v - E\{g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta}\} / E\{g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta}\}$, $\tilde{\Lambda}_{iv} = \Lambda_{iv} - E\{\Lambda_{iv} g_i^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta}\} / E\{g_i^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta}\}$, $A^{\otimes 2} = AA^T$ for any matrix A . We will estimate the two unknown functions $m_\mu(u)$ and $m_v(t)$ nonparametrically. For notational simplicity, we use the same kernel function and bandwidth for both nonparametric regressions.

2.3 Initial Estimate of the Mean Function

Methods for estimating the mean function $\mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)$ in (1)-(2) are already well-established if the potential heteroscedasticity is ignored, as we reviewed in Section 1. For given $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, $m_\mu(\cdot)$ is estimated via local linear regression of $Y - \mathbf{X}^T \boldsymbol{\beta}$ on $\mathbf{X}^T \boldsymbol{\alpha}$ resulting in $m_\mu(\hat{\cdot})$. More precisely, it should be $m_\mu(\hat{\cdot}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, but to keep the notation simpler we will sometimes use the convention $m_\mu(\hat{\cdot})$. Then $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are estimated by profiling. Thus, for any given $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, we use local linear regression to estimate m_μ by minimizing

$$\sum_{i=1}^n \left[Y_i - \{a + b(X_i^T \boldsymbol{\alpha} - u) + \mathbf{X}_i^T \boldsymbol{\beta}\} \right]^2 K_h(X_i^T \boldsymbol{\alpha} - u) \quad (4)$$

with respect to $\mathbf{a}$ and $\mathbf{b}$, where $\mathbf{K}$ is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$ and h is the bandwidth. Let $(\hat{a}, \hat{b})$ be the minimizer of (4) and set $m_\mu(\hat{u}) = \hat{a}$.

We then estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ subject to the constraints $\|\boldsymbol{\alpha}\| = 1$ and $\boldsymbol{\alpha}^T \boldsymbol{\beta} = 0$ by minimizing

$$\sum_{i=1}^n \left[Y_i - \{\hat{m}_\mu(X_i^T \boldsymbol{\alpha}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathbf{X}_i^T \boldsymbol{\beta}\} \right]^2, \quad (5)$$

with respect to $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are functions of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$. We define the

solution of (5) as $\hat{\omega} = (\hat{\boldsymbol{\alpha}}_{\setminus 1}^T, \hat{\boldsymbol{\beta}}_{\setminus 1}^T)^T$. Thus the final estimator for $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \{\boldsymbol{\alpha}(\hat{\omega}), \boldsymbol{\beta}(\hat{\omega})\}$. We use $\hat{\mathbf{J}}_\omega$ to denote $\mathbf{J}_\omega$ evaluated at $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ and in what follows use $\mathbf{J}_\omega$ for $\mathbf{J}_\omega$ evaluated at the true value $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \{\boldsymbol{\alpha}(\omega_0), \boldsymbol{\beta}(\omega_0)\}$.

More efficiency for estimating (m_μ, α, β) can be obtained via generalized least squares, which we will discuss in Section 2.5.

2.4 Initial Estimate of the Variance Function

Davidian and Carroll (1987) gave the general methodology and theory for variance function estimation in the parametric case. They distinguish between methods based on squared residuals and those based on absolute residuals, the former being more efficient if the regressions errors ε_i are normally distributed, but call this potential efficiency gain "tenuous" because it is less robust to outliers. Here we use absolute residuals, and follow a profiling approach analogous to that in Section 2.3. Of course, if one chooses to use squared residuals, the same algorithm described below applies, by replacing R_i by $\hat{R}_i^2$ and $g(\cdot)$ by $g^2(\cdot)$ in (6) and (7) below. The asymptotic distribution results for (θ, ζ) described in Theorem 3.2 below are also easily modified.

Define $R = |Y - \mu(\mathbf{X}, \alpha, \beta, m_\mu)|$ and $\hat{R} = |Y - \mu(\mathbf{X}, \hat{\alpha}, \hat{\beta}, \hat{m}_\mu)|$. Recall that $E(|\varepsilon|) = 1$. Then, approximately, $E(R|\mathbf{X}) \approx g\{v(\mathbf{X}, \theta, \zeta, m_\mu)\}$. Thus, write

$$\begin{aligned}\hat{R}_i &= |Y_i - \mu(\mathbf{X}_i, \hat{\alpha}, \hat{\beta}, \hat{m}_\mu)| = |Y_i - \{\hat{m}_\mu(\mathbf{X}_i^T \hat{\alpha}) + \mathbf{X}_i^T \hat{\beta}\}|, \\ \hat{S}_i &= \{\hat{m}_\mu(\mathbf{X}_i^T \hat{\alpha}) + \mathbf{X}_i^T \hat{\beta}\} - \{m_\mu(\mathbf{X}_i^T \alpha) + \mathbf{X}_i^T \beta\}.\end{aligned}$$

We first minimize in (a_0, a_1)

$$\sum_{i=1}^n K_h(\mathbf{X}_i^T \theta - t) \left[\hat{R}_i - g\{a_0 + a_1(\mathbf{X}_i^T \theta - t) + \mathbf{X}_i^T \zeta\} \right]^2, \quad (6)$$

resulting in $m_v(\cdot, \theta, \zeta)$. We then estimate (θ, ζ) by minimizing

$$\sum_{i=1}^n \left[\hat{R}_i - g\{\hat{m}_v(\mathbf{X}_i^T \theta, \theta, \zeta) + \mathbf{X}_i^T \zeta\} \right]^2, \quad (7)$$

subject to $\|\theta\| = 1$ and $\theta^T \zeta = 0$, calling these estimates $(\hat{\theta}, \hat{\zeta})$.

2.5 More Efficient Estimation of the Mean Function

We now investigate how we can more efficiently estimate (α, β) via generalized least squares. Using the method in Section 2.4, form the weights $1/\hat{g}_i^2$. Then estimate $m_\mu(\cdot)$ by replacing (4) by

$$\sum_{i=1}^n \left[Y_i - \{a + b(\mathbf{X}_i^T \alpha - u) + \mathbf{X}_i^T \beta\} \right]^2 K_h(\mathbf{X}_i^T \alpha - u) / \hat{g}_i^2, \quad (8)$$

We then update the estimates of (α, β) by the same process, but the least squares minimization are weighted with weights $\hat{g}_i^{-2}$. As in Davidian and Carroll (1987), it can be shown that there is no effect on the estimates of (α, β) due to estimating g_i , so only this

single step is required for first-order asymptotics. This is a well-known phenomenon, see for example Carroll and Ruppert (1988). Specifically, we estimate α and β by minimizing

$$\sum_{i=1}^n \left[Y_i - \left\{ \hat{m}_\mu \left(\mathbf{X}_i^T \alpha; \alpha, \beta \right) + \mathbf{X}_i^T \beta \right\} \right]^2 / \hat{g}_i^2, \quad (9)$$

with respect to α_1 and β_1 . We define the solution of (9) as $\hat{\omega}_{wls} = \left(\hat{\alpha}_{w,1}^T, \hat{\beta}_{w,1}^T \right)^T$. Thus the weighted generalized least squared estimator for (α, β) is

$$\left(\hat{\alpha}_{wls}, \hat{\beta}_{wls} \right) = \left\{ \alpha \left(\hat{\omega}_{wls} \right), \beta \left(\hat{\omega}_{wls} \right) \right\}.$$

2.6 Updated Estimation of the Variance Function

Our final variance function method is identical in computation to that of the initial estimator in Section 2.4 except that the absolute residuals R_i are the absolute residuals from the

weighted fit, so that $\hat{R} = |Y - \mu \left\{ \mathbf{X}, \hat{\alpha}_{WLS}, \hat{\beta}_{WLS}, \hat{m} \left(\mathbf{X}^T \hat{\alpha}_{WLS}, \hat{\alpha}_{WLS}, \hat{\beta}_{WLS} \right) \right\}|$. We denote the resulting estimators as $\hat{\vartheta}_{WLS}$ and $m_{\hat{\vartheta}, WLS}$.

2.7 Computing

There is a universal theme to the computation of our estimators, namely that all the methods are based on fitting general models of the form $g\{m_*(\mathbf{X}^T \mathbf{a}) + \mathbf{X}^T \mathbf{b}\}$ with $\|\mathbf{a}\| = 1$ and $\mathbf{a}^T \mathbf{b} = 0$, while at the same time incorporating weights. Thus, for example, in Section 2.4, one could easily carry the process further and add weights to improve estimation of the variance function, although even in parametric models this step is usually not done.

2.8 Estimation of $\text{var}(\epsilon)$

In Section 1, we made the identifiability constraint that $E(|\epsilon|) = 1$, so that $g\{v(\mathbf{X}, \theta, \zeta, m_v)\}$ is proportional to $\text{var}(Y|\mathbf{X})$. To estimate the actual variance function, we need to estimate $\sigma^2 = \text{var}(\epsilon)$. Since $E[Y - \{m_\mu(X^T \alpha; \alpha, \beta) + X^T \beta\}]^2 / g^2 = \sigma^2$, it is natural to define

$$\hat{\sigma}_n^2 = (n - 2p + 2)^{-1} \sum_{i=1}^n \left[Y_i - \left\{ \hat{m}_\mu \left(X_i^T \hat{\alpha}; \hat{\alpha}, \hat{\beta} \right) + X_i^T \hat{\beta} \right\} \right]^2 / \hat{g}_i^2.$$

Under the conditions of Theorem 3.1 and assuming that $E(\epsilon^4) < \infty$, a straightforward but

tedious manipulation shows that $n^{1/2} \left(\hat{\sigma}_n^2 - \sigma^2 \right) \rightarrow \text{Normal} \left[0, E \left\{ \left(\epsilon^2 - \sigma^2 \right)^2 \right\} \right]$.

3 Asymptotic Results

3.1 Mean Function Estimation

Theorem 3.1 Suppose that Assumptions A and B in the Appendix hold. Define

$Q_\omega = E \left\{ \left(\mathbf{J}_\omega^T \tilde{\Lambda}_\mu \right)^{\otimes 2} \right\}$. As $n \rightarrow \infty$, $nh^4 \rightarrow 0$, for the unweighted mean function estimates of Section 2.3,

$$\begin{aligned}
& \hat{m}_\mu(u, \hat{\alpha}) \\
& - m_\mu(u) = n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \alpha - u) \frac{\varepsilon_i}{f_\alpha(u)} \\
& - E(\Lambda_\mu^T | \mathbf{X}^T \alpha) \mathbf{J}_\omega(\hat{\omega} - \omega) \\
& + o_p(n^{-1/2}); n^{1/2} Q_\omega(\hat{\omega} - \omega) \\
& = n^{-1/2} \sum_{i=1}^n \varepsilon_i \mathbf{J}_\omega^T \hat{\Lambda}_{i\mu} + o_p(1).
\end{aligned}$$

Accordingly,

$$n^{1/2} \left\{ (\hat{\alpha} - \alpha)^T, (\hat{\beta} - \beta)^T \right\}^T \rightarrow \text{Normal} \left[0, E \left\{ \varepsilon^2 \mathbf{J}_\omega Q_\omega^{-1} E(g \mathbf{J}_\omega^T \tilde{\Lambda}_\mu)^{\otimes 2} Q_\omega^{-1} \mathbf{J}_\omega^T \right\} \right]. \text{ Further,}$$

for the weighted mean function estimates of Section 2.5, Define $Q_{w,\omega} = E \left\{ \left(\mathbf{J}_\omega^T \tilde{\Lambda}_\mu / g \right)^{\otimes 2} \right\}$. Then

$$\begin{aligned}
& \hat{m}_\mu(u, \hat{\alpha}_{wLS}) \\
& - m_\mu(u) = n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \alpha - u) \frac{\varepsilon_i}{f_\alpha(u) E(1/g^2 | \mathbf{X}^T \alpha)} \\
& - E(\Lambda_\mu^T / g^2 | \mathbf{X}^T \alpha) \\
& / E(1/g^2 | \mathbf{X}^T \alpha) \mathbf{J}_\omega(\hat{\omega}_{wLS} - \omega) \\
& + o_p(n^{-1/2}); n^{1/2} Q_{w,\omega}(\hat{\omega}_{wLS} - \omega) \\
& = n^{-1/2} \sum_{i=1}^n \varepsilon_i \mathbf{J}_\omega^T \tilde{\Lambda}_{i\mu} \\
& / g_i^2 + o_p(1).
\end{aligned} \tag{10}$$

$$\text{Accordingly, } n^{1/2} \left\{ (\hat{\alpha}_{wLS} - \alpha)^T, (\hat{\beta}_{wLS} - \beta)^T \right\}^T \rightarrow \text{Normal} \left\{ 0, E \left(\varepsilon^2 \mathbf{J}_\omega Q_{w,\omega}^{-1} \mathbf{J}_\omega^T \right) \right\}.$$

Further, when ε is normally distributed, these estimators are the most efficient in the sense of semiparametric efficiency (Bickel et al., 1993).

Remark 1 There are 3 implications of Theorem 3.1. First and most obvious, the weighted estimates $\hat{\omega}_{wLS}$ are more efficient than their unweighted version $\hat{\omega}$. Second, the function estimated $\hat{m}_\mu(u, \hat{\alpha}_{wLS})$ is also more efficient than its unweighted version $\hat{m}_\mu(u, \hat{\alpha})$. Finally, an implication of Theorem 3.1 is that $\hat{\omega}_{wLS}$ is asymptotically oracle, in the sense that it has the same limiting distribution as if the variance function were known. However, in smaller sample size situations, because of the nonparametric function estimation, we expect that knowing the true variance function will result in more efficient estimation of ω as well as $m_\mu(\cdot)$. All three points are confirmed in the simulation study of Section 4.2.

Remark 2 The limiting asymptotic distribution for $(\hat{\alpha}_{wLS}, \hat{\beta}_{wLS})$ described in Theorem 3.1 has an efficiency property even when ε is not normally distributed. The asymptotic

distribution described there has a seemingly different expression from the asymptotic variance from the estimators derived by Ma and Zhu (2013a), which are described by them as semiparametric efficient for the case that the variance function is known or estimated at a sufficiently fast rate. Ma and Zhu used a different parameterization from ours in order to achieve identifiability. However, it can be shown that the two asymptotic variances are the same if we use their parameterization.

3.2 Variance Function Estimation

Theorem 3.2 Suppose that Assumptions A and B in the Appendix hold. Define

$$Q_{\vartheta} = E \left\{ \mathbf{J}_{\vartheta}^T \tilde{\Lambda}_v g^{(1)} \right\}^{\otimes 2}. \text{ Define}$$

$$\Sigma_{\vartheta} = E \left[\varepsilon \left\{ \mathbf{J}_{\vartheta}^T E \left(g^{(1)} \Lambda_v d|U \right) + (\mathbf{J}_{\vartheta} - \mathbf{J}_{\omega})^T E \left(g^{(1)} \Lambda_v d\tilde{\Lambda}_{\mu} \right) \mathbf{J}_{\omega} Q_{\omega}^{-1} \mathbf{J}_{\omega}^T \tilde{\Lambda}_{\mu} \right\} + \delta \mathbf{J}_{\vartheta}^T \tilde{\Lambda}_v g^{(1)} \right]^{\otimes 2}; \quad (11)$$

$$\Sigma_{\vartheta, wls} = E \left[\varepsilon \left\{ \mathbf{J}_{\vartheta}^T E \left(g^{(1)} \Lambda_v d|U \right) + (\mathbf{J}_{\vartheta} - \mathbf{J}_{\omega})^T E \left(g^{(1)} \Lambda_v d\tilde{\Lambda}_{\mu} \right) \mathbf{J}_{\omega} Q_{w, \omega}^{-1} \mathbf{J}_{\omega}^T \tilde{\Lambda}_{\mu} / g^2 \right\} + \delta \mathbf{J}_{\vartheta}^T \tilde{\Lambda}_v g^{(1)} \right]^{\otimes 2}. \quad (12)$$

Then for the initial estimator of Section 2.4, as $n \rightarrow \infty$, $nh^4 \rightarrow \infty$ and $nh^6 \rightarrow 0$, dropping arguments for parameters,

$$\begin{aligned} \hat{m}_v(t) - m_v(t) &= n^{-1} \sum_{i=1}^n K_h \left(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t} \right) \frac{g_i^{(1)} \delta_i}{f_{\theta}(t) E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \\ &\quad - \left[\frac{E \{ g^{(1)2}(\cdot) \Lambda_v | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}}{E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \right]^T \mathbf{J}_{\vartheta} (\hat{\vartheta} - \vartheta) \\ &\quad - \left[\frac{E \{ g^{(1)}(\cdot) d\tilde{\Lambda}_{\mu} | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}}{E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \right]^T \mathbf{J}_{\omega} (\hat{\omega} - \omega) \\ &\quad + o_p \left(n^{-1/2} \right), \end{aligned} \quad (13)$$

For the updated variance function estimator of Section 2.6, again dropping arguments for parameters,

$$\begin{aligned} \hat{m}_{v, WLS}(t) &\quad - m_v(t) = n^{-1} \sum_{i=1}^n K_h \left(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t} \right) \frac{g_i^{(1)} \delta_i}{f_{\theta}(t) E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \\ &\quad - \left[\frac{E \{ g^{(1)2}(\cdot) \Lambda_v | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}}{E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \right]^T \mathbf{J}_{\vartheta} (\hat{\vartheta}_{WLS} - \vartheta) \\ &\quad - \left[\frac{E \{ g^{(1)}(\cdot) d\tilde{\Lambda}_{\mu} | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}}{E \{ g^{(1)2}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \}} \right]^T \mathbf{J}_{\omega} (\hat{\omega}_{WLS} - \omega) \\ &\quad + o_p \left(n^{-1/2} \right), \end{aligned} \quad (14)$$

Finally, for the parameter estimators,

$$n^{1/2}Q_{\vartheta}(\hat{\vartheta} - \vartheta) \rightarrow \text{Normal}(0, \Sigma_{\vartheta}); \quad (15)$$

$$n^{1/2}Q_{\vartheta}(\hat{\vartheta}_{WLS} - \vartheta) \rightarrow \text{Normal}(0, \Sigma_{\vartheta, WLS}). \quad (16)$$

Remark 3 Recalling that $d = \text{sign}(\varepsilon)$, in the special case that ε is symmetric, $E(d) = 0$ and the first two terms in the asymptotic covariance matrices (11)-(12) = 0, resulting in considerable simplification. Also, in this case, $\Sigma_{\vartheta} = \Sigma_{\vartheta, WLS}$. In addition, the last terms in (13)-(14) also equal 0. Indeed, in this case, the initial and updated variance function estimators are asymptotically equivalent to the estimator when the mean function is known.

In small sample size situations, however, because $(\hat{\alpha}_{WLS}, \hat{\beta}_{WLS}, \hat{m}_{WLS})$ are more efficient than their unweighted version, we expect them to have some effect on the variance estimators.

4 Numerical examples

4.1 Empirical Example

We use the NMMAPS (National Morbidity Mortality Air Pollution Study) database which contains daily mortality, weather and pollution data for 1987-2000. Here we only consider data for the year 1997. We will use the partially linear single-index model to explore the relationship between daily mean ozone level and some predictors. The selected 7 explanatory variables include mean temperature, relative humidity, mean CO_2 level, mean PM_{10} level, mean SO_2 level, daily humidity range, and daily temperature range. After excluding one day with missing observations, we have a sample size of $n = 364$. We used $g(v) = \exp(v)$. In Table 1 we display estimates and their standard errors for the second stage estimators, the latter computed following similar arguments as in Section 7 of Carroll *et al.* (1997).

There are some striking conclusions from the analysis. First, every parameter associated with the mean function is highly statistically significantly different from zero. In addition, all the single components of the single-index variance parameter, θ except mean SO_2 level are highly statistically significant. For the partially linear parameter, ζ , the coefficients associated with mean temperature, mean CO_2 level and mean SO_2 level are highly statistically significant. It is also clear that a complete partition of the mean or variance function parameters into the single-index component or the partially linear component is not supported by the analysis. Finally, it is obvious upon inspection, or by a formal hypothesis test, that $\alpha \neq \theta$ and that $\beta \neq \zeta$. Thus, the variance of Y given $\mathbf{X}$ is not a function of the mean.

Now consider the mean function. In the two left panels of Figure 3, we show the scatterplot of $Y_i - \mathbf{X}_i^T \hat{\beta}$ vs. $\mathbf{X}_i^T \hat{\alpha}$ for the first and second stage estimators, respectively. In both cases, the solid line is the fitted link $m_{\mu}(\cdot)$ for the mean function. It is evident that the means are not constant. There is a hint that $m_{\mu}(\cdot)$ might be constant for smaller values of $\mathbf{X}^T \alpha$, with a linear change afterwards, although a quadratic function might also give a reasonable fit.

Next consider the variance function. As stated above, we used $g(v) = \exp(v)$, so that one would expect that $\log(\hat{R})$ would be roughly linear when plotted against $\hat{m}_\nu(\cdot) + \mathbf{X}^T \hat{\zeta}$. In the two right panels of Figure 3, we show the scatterplot of $\log(\hat{R}_i)$ vs. $\hat{m}_\nu(\mathbf{X}_i^T \hat{\theta}) + \mathbf{X}_i^T \hat{\zeta}$ for the first and second stage estimators respectively, where the solid line is just the line that goes through the origin with slope one shown as a reference. Loess fits to these graphs are effectively linear except for some curvature caused by the large positive values along the x-axis, and a quadratic fit has no statistically significant quadratic term.

4.2 Simulations

We generated data from the model (1)-(3), with $g(x) = \exp(x)$ and $\varepsilon \sim \text{Normal}(0, \sigma^2)$. The covariates $\mathbf{X} = (X_1, \dots, X_8)^T$ are generated from a multivariate Gaussian distribution with covariance given by $\text{cov}(X_i, X_j) = (0.2)^{|i-j|}$. We set $\alpha = (1.0, 1.0, 1.0, 1.0, 0.5, 0.5, 0.5, 0.5)^T$, $\beta = (1.0, -1.0, 1.0, -1.0, 1.0, -1.0, 1.0, -1.0)^T$, $\theta = (0.5, -1, 0.5, -1, -0.5, -1, -1, -0.5)^T$ and $\zeta = (0.2, 0, -0.2, 0, 0.2, 0, -0.25, 0)^T$. As presented above, $\alpha^T \beta = \theta^T \zeta = 0$ but α and θ are not normalized to have unit norm, although α and θ are normalized for estimation.

The two examples differ in the functions m_μ and m_ν chosen.

- Example 1. $m_\mu(x) = 15 \sin(0.4x)$ and $m_\nu(x) = \cos(0.5x) + 3/2$.
- Example 2. $m_\mu(x) = x^2$ and $m_\nu(x) = 2/\{1 + \exp(-2x)\} + 3/2$.

We consider sample size $n = 200$ and three noise levels $\sigma = 0.2, 0.5, 1.0$. For each scenario, we generated 100 data sets and considered the estimators described in the outline in Section 2.1 and defined in Sections 2.3-2.6, along with two infeasible estimators.

- The unweighted least squares mean function estimates, Section 2.3.
- The initial variance function estimates starting from the unweighted least squares mean function estimates, Section 2.4.
- The weighted least squares mean function estimates, Section 2.5.
- The updated variance function estimates, Section 2.6.
- The infeasible weighted least squares estimates when the variance function is known, where "infeasible" means that these are not real estimators of practical utility since the variance functions are not known.
- The infeasible variance function estimates when the means are known.

The results are given in Tables 2-3. These tables display the root mean squared errors (RMSE) of different quantities, for the two examples respectively. For example, for α the RMSE is just $\|\hat{\alpha} - \alpha\|$ (the vectors are normalized to have unit norms) and for m_μ the RMSE

is $\sqrt{\sum_{t=1}^{50} \{\hat{m}_\mu(x_t) - m_\mu(x_t)\}^2 / 50}$, where $(x_1, \dots, x_{50})$ are equally-spaced grid points on an interval with lower bound the 0.01 quantile of the sampled $\mathbf{X}_i^T \alpha$ values and upper bound its 0.99 quantile. We see that σ has a significant effect on the estimation of the mean, but has a much smaller effect in the estimation of the variance, a known phenomenon in parametric problems, see Davidian and Carroll (1987).

We expect the mean function and its parameter estimates without weighting to be less efficient than that with weighting, and this is very clearly seen. Although asymptotically the weighted estimates are as efficient as the infeasible mean function and parameter estimates, there are small sample effects especially in Table 3 which confirm that the latter has more efficiency. Variances are generally much harder to estimate than means, and the small sample size effects are clearly seen, with the updated variance function and parameter estimates dominating the initial estimates, and not as efficient as the infeasible version.

In Figures 1 and 2, we show 20 randomly chosen estimated link functions for the two examples, where the true nonparametric link functions are shown as the solid red curves. We see that these link functions are estimated reasonably well.

Finally, we investigate the accuracy of the standard deviation estimates for the estimated parameters. To save space, we only consider the second-stage estimators. The true and estimated standard deviations (averaged over 100 data sets) are shown in Tables 4 and 5, for Example 1 and Example 2 respectively. The results reported in these tables indicate that the performance of the standard deviation estimation is reasonably accurate.

5 Discussion and Extensions

Most partially linear single index models (Carroll *et al.*, 1997; Liang *et al.*, 2010; Xia and Härdle, 2006) have a set of covariates dedicated to the parametric part, and a non-overlapping set of covariates dedicated to the nonparametric part. Our model (1)-(3) avoids partitioning the covariates, and makes the model more flexible and general. A priori partitioning can be done in our context, however, e.g., set $\alpha = (\alpha^T, \mathbf{0}^T)^T$ and $\beta = (\mathbf{0}^T, \beta^T)^T$, where α and β are nonzero vectors. This still leaves a gap between knowing a partition exactly and knowing nothing about the partition. Oracle penalization methods such as the adaptive Lasso and SCAD can be used to estimate the partition. Essentially, equation (5) is penalized in the usual way to get oracle estimates of (α, β) . Following this, (7) would have standard penalties for (θ, ζ) . Under standard conditions, such methods will be oracle, with the penalization parameters estimated by AIC or BIC. We will pursue the details of such an approach in the future.

There are many interesting possible extensions of our models (2)-(3). Most generally, our model implies that the distribution of Y given $\mathbf{X}$ depends on d ($= 4$ in our case) linear combinations of $\mathbf{X}$, where necessarily $d < p$. There is a vast literature on this general problem, mostly dealing with dimension reduction, see Ma and Zhu (2013b) for a recent example and many references. This general perspective is too general in practice when the main goal is to estimate a variance function, because one still needs to understand which linear combinations affect the variance function, which affect the mean function, and which, if any, affect both. A simpler model that still generalizes (2)-(3) is when $m_\mu(\cdot)$ and $m_v(\cdot)$ depend on d_μ and d_v linear combinations of $\mathbf{X}$, respectively. In the homoscedastic case, the problem of estimating the mean function in this resulting multiple-index model has been discussed by Xia (2008). Our estimation approach in Sections 2.3-2.6 can in principle be readily adapted to this more general problem, depending on how one estimates the linear combinations. It would be interesting to find limiting distributions and discuss efficiency in

this case. However, there are considerable issues with dimensionality, multivariate kernel functions, and bandwidth selection. We agree with the sentiment expressed in remark 5.3 of Xia (2008) that because of these issues, d_μ and d_v should be small, and our case that $d_\mu = d_v = 1$ is (the most in our opinion) "appealing" from a practical perspective.

Semiparametric asymptotic efficiency for estimating $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ is a much more complex problem, both technically and practically. We established in Theorem 3.2 that estimation of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)$ has no effect on estimation of the variance parameters $(\boldsymbol{\theta}, \boldsymbol{\zeta})$. Thus, from a theoretical perspective, working on the residuals from the mean fit is equivalent to a model that we observe $R = |Y - m_\mu(\mathbf{X}^T \boldsymbol{\alpha}) - \mathbf{X}^T \boldsymbol{\beta}|$, and working with R , the model can be thought of as

$$E(R|\mathbf{X}) = g \left\{ m_v(\mathbf{X}^T \boldsymbol{\theta}) + \mathbf{X}^T \boldsymbol{\zeta} \right\}; \quad (17)$$

$$\text{var}(R|\mathbf{X}) = \kappa^2 g^2 \left\{ m_v(\mathbf{X}^T \boldsymbol{\theta}) + \mathbf{X}^T \boldsymbol{\zeta} \right\}, \quad (18)$$

where κ is unknown. This is a case where the variance is proportional to a known function of the mean, a semiparametric problem that has not been solved in the literature. Indeed, it is quite a tricky problem, since R is not normally distributed

From a practical perspective, it is not obvious that this is a problem that should be solved from an efficient semiparametric perspective, at least in any practical context, for the following reasons.

- Except for the nonparametric flavor due to $m_v(\cdot)$, models (17)-(18) can be thought of as generalized linear models. The overwhelming practice in statistics for such models is to use the variance function only for weighting, but not to try to exploit it to obtain further asymptotic efficiency. It is easy to implement this weighting, and derive an analogue of Theorem 3.2. We have done so, but the promised asymptotic gains in efficiency for estimating $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ were not realized in our simulations.
- In especially the third author's fairly extensive experience with parametric variance function models (Carroll and Ruppert, 1982, 1988), the more that higher order moments are involved in the score functions, the worse the practical performance, and the longer it takes to reach asymptotics, as in our simulations described in the previous point.
- Our methods are practical, theoretically justified, and clearly work well in simulations and examples.

Acknowledgements

We thank the associate editor and two anonymous reviewers for their insightful comments that improved the manuscript significantly. We also thank Dr. Yingcun Xia for providing us with computational code. Lian's research was supported by a Singapore MOE Tier 1 grant. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444 and by Award Number 11228103, given by National Natural Science Foundation of China. Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030).

Appendix

A.1 Assumptions

Assumptions A:

- A1** The density function $f_{\alpha}(u)$ of $\mathbf{X}^T \boldsymbol{\alpha}$ is bounded away from zero and continuously differentiable on $(\mathbf{X}^T \boldsymbol{\alpha}, \mathbf{X} \in \Omega_X)$ and Ω_X is the support of $\mathbf{X}$ assumed to be compact. Furthermore, $f_{\alpha}(u)$ is uniformly continuous for $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$.
- A2** The function $m_{\mu}(\cdot)$ is twice continuously differentiable.
- A3** The kernel K is a bounded and symmetric probability density function, satisfying

$$\int_{-\infty}^{\infty} u^2 K(u) du \neq 0, \quad \int_{-\infty}^{\infty} |u|^l K(u) du < \infty, \quad l=1, 2, 3.$$

- A4** The matrix Q_{ω} is positive definite.

Assumptions B:

- B1** The density function $f_{\theta}(t)$ of $\mathbf{X}^T \boldsymbol{\theta}$ is bounded away from zero and continuously differentiable on $(\mathbf{X}^T \boldsymbol{\theta}, \mathbf{X} \in \Omega_X)$. Furthermore, $f_{\theta}(t)$ is uniformly continuous for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$.
- B2** The functions $m_{\nu}(\cdot)$ and $g(\cdot)$ are twice continuously differentiable.
- B3** $E\{g^{(1)}(\cdot) d\tilde{\Lambda}_{\mu} | \mathbf{X}^T \boldsymbol{\theta} = t\}$, $E\{g^{(1)}(\cdot) | \mathbf{X}^T \boldsymbol{\theta} = t\}$, and $E\{g^{(1)}(\cdot) \Lambda_{\mu} | \mathbf{X}^T \boldsymbol{\theta} = t\}$ are continuous functions.
- B4** The matrix $\Sigma_{\mathcal{J}}$ defined at (11) is positive definite.

A.2 Proof of Theorem 3.1 for the Unweighted Estimators

The results in Theorem 3.1 for the unweighted estimator follows easily using the same arguments as in the proof of Theorem 4 in Carroll *et al.* (1997).

A.3 Proof of (13) and (15) in Theorem 3.2

We first give an expression for $\hat{R}_i - R_i$. A direct simplification yields that

$$\begin{aligned} \hat{S}_i &= \left\{ \hat{m}_{\mu}(\mathbf{X}_i^T \hat{\boldsymbol{\alpha}}) + \mathbf{X}_i^T \hat{\boldsymbol{\beta}} \right\} - \left\{ m_{\mu}(\mathbf{X}_i^T \boldsymbol{\alpha}) + \mathbf{X}_i^T \boldsymbol{\beta} \right\} \\ &= \hat{m}_{\mu}^{(1)}(\mathbf{X}_i^T \boldsymbol{\alpha}) \mathbf{X}_i^T (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \mathbf{X}_i^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \hat{m}_{\mu}(\mathbf{X}_i^T \boldsymbol{\alpha}) - m_{\mu}(\mathbf{X}_i^T \boldsymbol{\alpha}) + o_p(n^{-1/2}) \\ &= \tilde{\Lambda}_{i\mu}^T \mathbf{J}_{\omega} (\hat{\omega} - \omega) + n^{-1} \sum_{j=1}^n K_h(U_j - U_i) \frac{\varepsilon_j}{f_{\alpha}(U_i)} + o_p(n^{-1/2}). \end{aligned}$$

From an identity (Knight, 1998, pp.758), we have that

$$\begin{aligned}
\hat{R}_i - R_i &= |Y_i - \{\hat{m}_\mu(\mathbf{X}_i^T \hat{\boldsymbol{\alpha}}) + \mathbf{X}_i^T \hat{\boldsymbol{\beta}}\}| - |Y_i - \{m_\mu(\mathbf{X}_i^T \boldsymbol{\alpha}) + \mathbf{X}_i^T \boldsymbol{\beta}\}| \\
&= -\hat{S}_i \left\{ I_{(\varepsilon_i > 0)} - I_{(\varepsilon_i \leq 0)} \right\} + \int_0^{\hat{S}_i} \left\{ I_{(\varepsilon_i < s)} - I_{(\varepsilon_i \leq 0)} \right\} ds \\
&= -d_i \tilde{\Lambda}_{i\mu}^T \mathbf{J}_\omega (\hat{\omega} - \omega) + d_i n^{-1} \sum_{j=1}^n K_h(U_j - U_i) \frac{\varepsilon_j}{f_\alpha(U_i)} + \int_0^{\hat{S}_i} \left\{ I_{(\varepsilon_i < s)} - I_{(\varepsilon_i \leq 0)} \right\} ds + o_p(n^{-1/2}).
\end{aligned} \tag{A.1}$$

Write $\iota_k = \int s^k K^2(s) ds$ and $\kappa_k = \int s^k K(s) ds$ for $k = 0, 1, 2$.

Step 1. Minimization of (6) is equivalent to $(\hat{a}_0, \hat{a}_1)$ being the solution of the equation:

$$\begin{aligned}
&n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left[\hat{R}_i - g \left\{ \hat{a}_0 + \hat{a}_1 (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\} \right] \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T \\
&\quad \times g^{(1)} \left\{ \hat{a}_0 + \hat{a}_1 (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\} = 0.
\end{aligned}$$

Using Taylor expansion and the assumptions on h , we know that the left hand side is

$$\begin{aligned}
&n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) (\hat{R}_i - R_i) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T g_i^{(1)} \\
&\quad + n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) (R_i - g_i) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T g_i^{(1)} \\
&\quad + n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left[g_i - g \left\{ \hat{a}_0 + \hat{a}_1 (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\} \right] \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T \\
&\quad \times g^{(1)} \left\{ \hat{a}_0 + \hat{a}_1 (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\} \\
&\quad + o(n^{-1/2}).
\end{aligned} \tag{A.2}$$

It follows from (A.1) that the first term in (A.2) is

$$\begin{aligned}
&-n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T g_i^{(1)} d_i \tilde{\Lambda}_{i\mu}^T \mathbf{J}_\omega (\hat{\omega} - \omega) \\
&\quad + n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T g_i^{(1)} d_i n^{-1} \sum_{j=1}^n K_h(U_j - U_i) \varepsilon_j \\
&\quad / f_\alpha(U_i) + n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t) / h \right\}^T g_i^{(1)} \int_0^{\hat{S}_i} \left\{ I_{(\varepsilon_i < s)} - I_{(\varepsilon_i \leq 0)} \right\} ds \\
&\quad + o_p(n^{-1/2}).
\end{aligned} \tag{A.3}$$

For the second term in (A.3), it's obvious that we can replace $\hat{\theta}$ by θ . Also, the term

$g_i^{(1)} / f_\alpha(U_i)$ is a nuisance and we eliminate it. We only consider the first component in the 2×1 vector, since the other one is similarly $o_p(n^{-1/2})$. In this case, with $d_i = \text{sign}(\varepsilon_i)$, we want to analyze

$$A_n = n^{-2} \sum_{i=1}^n \sum_{j=1}^n d_i \epsilon_j K_h(U_i - t) K_h(U_j - U_i).$$

Since $E(\epsilon_i) = 0$, $E(A_n) = 0$. We merely need to show that $\text{var}(A_n) = o(n^{-1})$. However,

$$\text{var}(A_n) = n^{-4} \sum_{i,j,p=1}^n \sum_{\ell \neq j} E \{ d_i \epsilon_j d_p \epsilon_\ell K_h(U_i - t) K_h(U_j - U_i) K_h(U_p - t) K_h(U_\ell - U_p) \}.$$

All the various combinations can be done in turn. For example, if $i=j=p=\ell$, then the contribution to $\text{var}(A_n)$ is $O\{(nh)^{-3}\}$, so we only need that $(nh)^{-3} = o(n^{-1})$. If any 3 of the (i, j, p, ℓ) are equal but different from the remaining one, the contribution is zero. If all 4 are unequal, the contribution is 0. So, the only cases we need to copy with are those in which $(i=j, p=\ell)$, $(i=p, j=\ell)$ and $(i=\ell, j=p)$. All these cases are similar, so consider the case that $i = p$ and $j = \ell$, and write $c = E(\epsilon_i^2) E(d_j^2)$. Then the contribution to $\text{var}(A_n)$ is

$$\begin{aligned} B_n &= cn^{-2} E \{ K_h^2(U_1 - t) K_h^2(U_2 - U_1) \} \\ &= cn^{-2} h^{-4} \int K_h^2(u_1 - t) K_h^2(u_2 - u_1) f_U(u_1) f_U(u_2) du_1 du_2. \end{aligned}$$

Make the change of variables $z_1 = (u_1 - t)/h$ and $z_2 = (u_2 - u_1)/h$, so that

$$B_n = c(nh)^{-2} \int K^2(z_1) K^2(z_2) f_U(t - z_1 h) f_U(t - z_2 h) dz_1 dz_2 = O\{(nh)^{-2}\}.$$

This means that $\text{var}(A_n) = o(n^{-1})$ as long as $nh^2 \rightarrow \infty$, which follows from our Assumptions. Analogously we can prove that the third term in (A.3) is also of order $(n^{-1/2})$.

The second term in (A.2) equals $n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t)/h \right\}^T g_i^{(1)} \delta_i$, while the third term in (A.2) can be further expressed as

$$\begin{aligned} & -n^{-1} \sum_{i=1}^n \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t)/h \right\}^T \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t)/h \right\} K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) g_i^{(1)2} (\hat{a}_0 - a_0, \hat{a}_1 - a_1)^T \\ & - n^{-1} \sum_{i=1}^n K_h(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t}) \left\{ 1, (\mathbf{X}_i^T \hat{\boldsymbol{\theta}} - t)/h \right\}^T g_i^{(1)2} \Lambda_{iv}^T \mathbf{J}_\vartheta (\hat{\vartheta} - \vartheta) \\ & = -E \left\{ g^{(1)2} | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \right\} f_\theta(t) \text{diag}(1, \kappa_2) (\hat{a}_0 - a_0, \hat{a}_1 - a_1)^T \\ & - (1, 0)^T E \left\{ g^{(1)2} \Lambda_v^T | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \right\} f_\theta(t) \mathbf{J}_\vartheta (\hat{\vartheta} - \vartheta) \\ & + o_p(n^{-1/2}). \end{aligned}$$

A combination of these arguments yields that

$$\begin{aligned}
& E \left\{ g^{(1)2} | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \right\} f_{\theta}(t) (\hat{a}_0 - a_0) \\
& + E \left\{ g^{(1)2} \Lambda_v^T | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \right\} f_{\theta}(t) \mathbf{J}_{\vartheta} (\hat{\vartheta} - \vartheta) \\
& + f_{\theta}(t) E \left\{ g^{(1)} d\tilde{\Lambda}_{\mu}^T | \mathbf{X}^T \boldsymbol{\theta} = \mathbf{t} \right\} \mathbf{J}_{\omega} (\hat{\omega} - \omega) \\
& - n^{-1} \sum_{i=1}^n K_h \left(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{t} \right) g_i^{(1)} \delta_i \\
& = o_p \left(n^{-1/2} \right),
\end{aligned}$$

from which (13) follows.

Step 2. Set $\hat{g}_i := g \left\{ \hat{m}_v \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}} \right) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\}$, $\hat{g}_i^{(1)} := g^{(1)} \left\{ \hat{m}_v \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}} \right) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\}$, and

let $\hat{\Lambda}_{iv}$ be Λ_{iv} evaluated at $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}})$. Minimization of (7) is equivalent to solving

$\sum_{i=1}^n \mathbf{J}_{\vartheta}^T \hat{\Lambda}_{iv} \left(\hat{R}_i - \hat{g}_i \right) \hat{g}_i^{(1)} = 0$ Now, $\sum_{i=1}^n \mathbf{J}_{\vartheta}^T \hat{\Lambda}_{iv} \left(\hat{R}_i - \hat{g}_i \right) \hat{g}_i^{(1)}$ is

$$\begin{aligned}
& \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - \hat{g}_i \right) \hat{g}_i^{(1)} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(\hat{R}_i - \hat{g}_i \right) \hat{g}_i^{(1)} + o_p \left(n^{1/2} \right) \\
& = \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - R_i \right) \hat{g}_i^{(1)} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(R_i - \hat{g}_i \right) \hat{g}_i^{(1)} \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \lambda_{iv} \left(\hat{R}_i - \hat{g}_i \right) \hat{g}_i^{(1)} + o_p \left(n^{1/2} \right) \\
& = \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - R_i \right) g_i^{(1)} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - R_i \right) \left(\hat{g}_i^{(1)} - g_i^{(1)} \right) \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(R_i - g_i \right) \hat{g}_i^{(1)} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(g_i - \hat{g}_i \right) \hat{g}_i^{(1)} \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(R_i - \hat{g}_i \right) \hat{g}_i^{(1)} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(\hat{R}_i - R_i \right) \hat{g}_i^{(1)} + o_p \left(n^{1/2} \right).
\end{aligned}$$

This is the sum of 10 terms given by

$$\begin{aligned}
& \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - R_i \right) g_i^{(1)} \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{R}_i - R_i \right) \left(\hat{g}_i^{(1)} - g_i^{(1)} \right) \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(R_i - g_i \right) \hat{g}_i^{(1)} \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \left(\hat{\Lambda}_{iv} - \Lambda_{iv} \right) \left(\hat{g}_i - g_i \right) \hat{g}_i^{(1)} \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(\hat{R}_i - R_i \right) \left(\hat{g}_i^{(1)} - g_i^{(1)} \right) \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(R_i - g_i \right) \left(\hat{g}_i^{(1)} - g_i^{(1)} \right) \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(g_i - \hat{g}_i \right) \left(\hat{g}_i^{(1)} - g_i^{(1)} \right) \\
& + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(\hat{R}_i - R_i \right) g_i^{(1)} \\
& - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \left(\hat{g}_i - g_i \right) g_i^{(1)} \\
& + o_p \left(n^{1/2} \right) \triangleq \sum_{k=1}^{10} I_k \\
& + o_p \left(n^{1/2} \right).
\end{aligned} \tag{A.4}$$

Under our assumptions, the first 7 terms are easily seen to be $o_p(n^{1/2})$. In addition,

$$\begin{aligned}
l_{10} &= - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} g_i^{(1)} \left[g \left\{ \hat{m}_v \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}} \right) + \mathbf{X}_i^T \hat{\boldsymbol{\zeta}} \right\} - g \left\{ m_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) + \mathbf{X}_i^T \boldsymbol{\zeta} \right\} \right] \\
&= - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} g_i^{(1)2} \left[\left\{ \hat{m}_v \left(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}} \right) - \hat{m}_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) \right\} + \mathbf{X}_i^T \left(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta} \right) + \left\{ \hat{m}_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) - m_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) \right\} \right] + o \left(n^{1/2} \right).
\end{aligned}$$

Now, we use the (13) for $\left\{ \hat{m}_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) - m_v \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) \right\}$ and obtain that

$$\begin{aligned}
I_{10} &= - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \Lambda_{iv}^T \mathbf{J}_{\vartheta} g_i^{(1)2} \left(\hat{\vartheta} - \vartheta \right) - \sum_{i=1}^n \frac{\mathbf{J}_{\vartheta}^T \Lambda_{iv} g_i^{(1)2}}{f_{\vartheta} \left(\mathbf{X}_i^T \boldsymbol{\theta} \right) E \left\{ g^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} n^{-1} \sum_{j=1}^n K_h \left(\mathbf{X}_i^T \boldsymbol{\theta} - \mathbf{X}_j^T \boldsymbol{\theta} \right) g_j^{(1)} \delta_j + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} g_i^{(1)2} \left[\frac{E \left\{ g^{(1)2}(\cdot) \Lambda_v | \mathbf{X}_i^T \boldsymbol{\theta} \right\}}{E \left\{ g^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} \right]^T \\
&= - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \tilde{\Lambda}_{iv}^T \mathbf{J}_{\vartheta} g_i^{(1)2} \left(\hat{\vartheta} - \vartheta \right) - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \delta_i g_i^{(1)} \frac{E \left\{ \Lambda_{iv} g_i^{(1)2} | \mathbf{X}_i^T \boldsymbol{\theta} \right\}}{E \left\{ g^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} + \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} g_i^{(1)2} \left[\frac{E \left\{ g^{(1)}(\cdot) d \tilde{\Lambda}_{\mu} | \mathbf{X}_i^T \boldsymbol{\theta} \right\}}{E \left\{ g^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} \right]^T \mathbf{J}_{\omega} \left(\hat{\omega} - \omega \right) + o_p \left(n^{1/2} \right) \\
&= - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \Lambda_{iv} \tilde{\Lambda}_{iv}^T \mathbf{J}_{\vartheta} g_i^{(1)2} \left(\hat{\vartheta} - \vartheta \right) - \sum_{i=1}^n \mathbf{J}_{\vartheta}^T \delta_i g_i^{(1)} \frac{E \left\{ \Lambda_{iv} g_i^{(1)2} | \mathbf{X}_i^T \boldsymbol{\theta} \right\}}{E \left\{ g^{(1)2}(\cdot) | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} + n \mathbf{J}_{\vartheta}^T E \left(\Lambda_v g^{(1)} d \tilde{\Lambda}_{\mu} \right) \mathbf{J}_{\omega} \left(\hat{\omega} - \omega \right) + o \left(n^{1/2} \right).
\end{aligned}$$

We now deal with the term I_8 . We have

$$\begin{aligned}
 & n^{-1} \sum_{i=1}^n \mathbf{J}_{\omega}^T \mathbf{\Lambda}_{iv} \left(\hat{R}_i - R_i \right) g_i^{(1)} \\
 & = -n^{-1} \sum_{i=1}^n g_i^{(1)} \mathbf{J}_{\omega}^T \mathbf{\Lambda}_{iv} d_i \tilde{\mathbf{\Lambda}}_{i\mu}^T \mathbf{J}_{\omega} (\hat{\omega} - \omega) \\
 & + n^{-2} \sum_{i,j} K_h(U_j - U_i) \varepsilon_j g_i^{(1)} \mathbf{J}_{\omega}^T \mathbf{\Lambda}_{iv} d_i \\
 & / f_{\alpha}(U_i) + n^{-1} \sum_{i=1}^n g_i^{(1)} \mathbf{J}_{\omega}^T \mathbf{\Lambda}_{iv} \int_0^{\hat{S}_i} \left\{ I_{\{\varepsilon \leq s\}} - I_{\{\varepsilon \leq 0\}} \right\} ds.
 \end{aligned} \tag{A.6}$$

For the 2nd term in (A.6), we consider $\mathbf{S}_n = n^{-1/2} \sum_{i=1}^n Q_i d_i n^{-1} \sum_{j=1}^n K_h(U_j - U_i) \varepsilon_j$, where $Q_i = g_i^{(1)} \mathbf{J}_{\theta}^T \mathbf{\Lambda}_{iv} / f_{\alpha}(U_i)$. We have that $E(\varepsilon_i) = 0$. Obviously, $\mathbf{S}_n = n^{-1/2} \sum_{i=1}^n \varepsilon_i n^{-1} \sum_{j=1}^n K_h(U_j - U_i) Q_j d_j$. We have that

$$\mathbf{S}_n = n^{-1/2} \sum_{i=1}^n \varepsilon_i f_{\alpha}(U_i) E(Qd|U_i) + o_p(1) = O_p(1). \tag{A.7}$$

It suffices to show that

$$\mathbf{T}_n = n^{-1/2} \sum_{i=1}^n \varepsilon_i \left\{ n^{-1} \sum_{j=1}^n K_h(U_j - U_i) Q_j d_j - f_{\alpha}(U_i) E(Qd|U_i) \right\} = o_p(1).$$

The mean of $\mathbf{T}_n = 0$ and the variance is $o_p(1)$. Here are the calculations. Write $\text{var}(\varepsilon_i) = G_i$

$$\begin{aligned}
 \text{var}(\mathbf{T}_n) &= n^{-3} \sum_{i=1}^n \sum_{k=1}^n \sum_{\ell=1}^n E[\varepsilon_i \varepsilon_k \{K_h(U_j - U_i) Q_j d_j - f_{\alpha}(U_i) E(Qd|U_i)\} \times \{K_h(U_{\ell} - U_k) Q_{\ell} d_{\ell} - f_{\alpha}(U_k) E(Qd|U_k)\}] \\
 &= n^{-3} \sum_{i=1}^n \sum_{j=1}^n \sum_{\ell=1}^n E[G_i \{K_h(U_j - U_i) Q_j d_j - f_{\alpha}(U_i) E(Qd|U_i)\} \times \{K_h(U_{\ell} - U_i) Q_{\ell} d_{\ell} - f_{\alpha}(U_i) E(Qd|U_i)\}].
 \end{aligned}$$

When $j = k$,

$$E[G_i \{K_h(U_j - U_i) Q_j d_j - f_{\alpha}(U_i) E(Qd|U_i)\} \times \{K_h(U_{\ell} - U_i) Q_{\ell} d_{\ell} - f_{\alpha}(U_i) E(Qd|U_i)\}] = O(h^4) = o(1).$$

Hence,

$$\begin{aligned}
 \text{var}(\mathbf{T}_n) &= n^{-3} \sum_{i=1}^n \sum_{j=1}^n E[G_i \{K_h(U_j - U_i) Q_j d_j - f_{\alpha}(U_i) E(Qd|U_i)\}] + o(1) \\
 &= O(n^{-1}) = o(1).
 \end{aligned}$$

For the 3rd term in (A.6), $\hat{S}_i = O_p((nh)^{-1/2})$. For any constant $C > 0$ (similar for $C < 0$), by a direct calculation of the mean and variance, we have

$$\begin{aligned}
\left| \int_0^{C/\sqrt{nh}} \left\{ I_{\{\varepsilon \leq s\}} - I_{\{\varepsilon \leq 0\}} \right\} ds \right| &= \left| E \int_0^{C/\sqrt{nh}} \left\{ I_{\{\varepsilon \leq s\}} - I_{\{\varepsilon \leq 0\}} \right\} ds \right| \{1 + o_p(1)\} \\
&= \left[E \int_0^{C/\sqrt{nh}} \{F(s) - F(0)\} ds \right] \{1 + o_p(1)\} \\
&= f'(0) C^2 / (2nh) \{1 + o_p(1)\} \\
&= O_p\{1/(nh)\} = o_p(n^{-1/2}),
\end{aligned}$$

where F is the conditional distribution function of ε and f is the corresponding density. This shows the 3rd term of (A.6) is also $o_p(n^{-1/2})$.

As a consequence,

$$I_8 = -n \mathbf{J}_\omega^T E \left(g^{(1)} \Lambda_v d\tilde{\Lambda}_\mu^T \right) \mathbf{J}_\omega (\hat{\omega} - \omega) + \sum_{i=1}^n \varepsilon_i f_\alpha(U_i) E(Qd|U_i) + o_p(n^{1/2}). \quad (\text{A.8})$$

A combination of (A.5), (A.8), and the expression given in (A.4) yields that

$$\begin{aligned}
&-n^{1/2} \mathbf{J}_\omega^T E \left(g^{(1)} \Lambda_v d\tilde{\Lambda}_\mu^T \right) \mathbf{J}_\omega (\hat{\omega} - \omega) \\
&+ n^{-1/2} \sum_{i=1}^n \varepsilon_i f_\alpha(U_i) E(Qd|U_i) \\
&+ n^{-1/2} \sum_{i=1}^n \mathbf{J}_\vartheta^T \Lambda_{iv} g_i^{(1)} \delta_i \\
&- n^{-1/2} \sum_{i=1}^n \mathbf{J}_\vartheta^T \Lambda_{iv} \tilde{\Lambda}_{iv}^T \mathbf{J}_\vartheta g_i^{(1)2} (\hat{\vartheta} - \vartheta) \\
&- n^{-1/2} \sum_{i=1}^n \delta_i g_i^{(1)} \mathbf{J}_\vartheta^T \frac{E \left\{ \Lambda_{iv} g_i^{(1)2} | \mathbf{X}_i^T \boldsymbol{\theta} \right\}}{E \left\{ g_i^{(1)2} | \mathbf{X}_i^T \boldsymbol{\theta} \right\}} \\
&+ n^{1/2} \mathbf{J}_\vartheta^T E \left(\Lambda_v g^{(1)} d\tilde{\Lambda}_\mu \right) \mathbf{J}_\omega (\hat{\omega} - \omega) \\
&+ o_p(1) = n^{-1/2} \sum_{i=1}^n \varepsilon_i f_\alpha(U_i) E(Qd|U_i) \\
&+ n^{-1/2} \sum_{i=1}^n \mathbf{J}_\vartheta^T \tilde{\Lambda}_{iv} g_i^{(1)} \delta_i \\
&- n^{1/2} Q_\vartheta (\hat{\vartheta} - \vartheta) \\
&+ (\mathbf{J}_\vartheta - \mathbf{J}_\omega)^T E(\mathbf{J}_\vartheta - \mathbf{J}_\omega)^T E \left(\Lambda_v g^{(1)} d\tilde{\Lambda}_\mu \right) \mathbf{J}_\omega Q_\omega^{-1} \sum_{i=1}^n \varepsilon_i \mathbf{J}_\omega^T \tilde{\Lambda}_{i\mu} + o_p(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
&n^{1/2} Q_\vartheta (\hat{\vartheta} - \vartheta) \\
&= n^{-1/2} \sum_{i=1}^n \left[\varepsilon_i \left\{ \mathbf{J}_\vartheta^T E \left(g^{(1)} \Lambda_v d\tilde{\Lambda}_\mu \right) + (\mathbf{J}_\vartheta - \mathbf{J}_\omega)^T E \left(\Lambda_v g^{(1)} d\tilde{\Lambda}_\mu \right) \mathbf{J}_\omega Q_\omega^{-1} \mathbf{J}_\omega^T \tilde{\Lambda}_{i\mu} \right\} + \delta_i \mathbf{J}_\vartheta^T \tilde{\Lambda}_{iv} g_i^{(1)} \right] + o_p(1),
\end{aligned}$$

completing the proof.

A.4 Proof of Theorem 3.1 for the Weighted Estimators

For the weighted estimators, under the conditions of Theorem 3.1, we follow similar arguments to those used by Ichimura (1993) and show that $\hat{\omega}$ is a root- n consistent estimator of ω . Because the proof is straightforward, we do not present it here. We next demonstrate the asymptotic normality of $\hat{\omega}$ by using a general result of Newey (1994).

Let $\mathcal{H}(U) = E(\mathbf{X}/g^2|U) / E(1/g^2|U)$, $\kappa = m_\mu^{(1)}(U) \{\mathbf{X} - \mathcal{H}(U)\}$. In addition, let

$$\Psi(\mathcal{H}, m_\mu, \kappa, g; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}) = \mathbf{J}_\omega \left(Y - m_\mu - \mathbf{X}^\top \boldsymbol{\beta} \right) \left\{ \kappa, \mathbf{X}^\top - \mathcal{H}^\top(U) \right\} / g^2.$$

For any given $\mathcal{H}^*$, m_μ^* , g^* , and κ^* , define

$$\begin{aligned} D \left(\mathcal{H}^* - \mathcal{H}, m_\mu^* - m_\mu, \kappa^* - \kappa, g^* - g, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \right) \\ = \frac{\partial \Psi}{\partial \mathcal{H}} (\mathcal{H}^* - \mathcal{H}) + \frac{\partial \Psi}{\partial m_\mu} (m_\mu^* - m_\mu) + \frac{\partial \Psi}{\partial \kappa} (\kappa^* - \kappa) + \frac{\partial \Psi}{\partial g} (g^* - g), \end{aligned}$$

where the partial derivatives are the Fréchet partial derivatives. We have

$$\begin{aligned} \partial \Psi / \partial \mathcal{H} &= \mathbf{J}_\omega \left(Y - m_\mu - \mathbf{X}^\top \boldsymbol{\beta} \right) (0, -1)^\top / g^2, \\ \partial \Psi / \partial m_\mu &= -\mathbf{J}_\omega \left\{ \kappa, \mathbf{X}^\top - \mathcal{H}^\top(U) \right\}^\top / g^2, \\ \partial \Psi / \partial \kappa &= \mathbf{J}_\omega \left(Y - m_\mu - \mathbf{X}^\top \boldsymbol{\beta} \right) (1, 0)^\top / g^2, \\ \partial \Psi / \partial g &= -2\mathbf{J}_\omega \left(Y - m_\mu - \mathbf{X}^\top \boldsymbol{\beta} \right) \left\{ \kappa, \mathbf{X}^\top - \mathcal{H}^\top(U) \right\}^\top / g^3. \end{aligned}$$

It is easy to verify that the expectations of these partial derivatives are zero. Accordingly,

$$\begin{aligned} & \left\| \Psi \left(m_\mu^*, m_\mu^*, \kappa^*, g^*; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \right) \right. \\ & \quad - \Psi \left(\mathcal{H}, m_\mu, \kappa, g; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \right) \\ & \quad \left. - D \left(\mathcal{H}^* - \mathcal{H}, m_\mu^* - m_\mu, \kappa^* - \kappa, g^* - g; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \right) \right\| \quad (\text{A.9}) \\ & = O \left(\left\| \mathcal{H}^* - \mathcal{H} \right\|^2 + \left\| m_\mu^* - m_\mu \right\|^2 + \left\| \kappa^* - \kappa \right\|^2 + \left\| g^* - g \right\|^2 \right), \end{aligned}$$

where $\|\cdot\|$ denotes the Sobolev norm, i.e., the supremum norm of the function itself as well as its derivatives. Equation (A.9) is Newey's assumption 5.1(i). It is also noteworthy that his assumption 5.2 holds by the expression of $D(\cdot, \cdot, \cdot; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X})$. Moreover, the result

$$E \left\{ D \left(\mathcal{H}^* - \mathcal{H}, m_\mu^* - m_\mu, \kappa^* - \kappa, g^* - g; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X} \right) \right\} = 0$$

leads to Newey's assumption 5.3.

Applying similar techniques to those used in Mack and Silverman (1982), we obtain the following expressions, which hold uniformly in $u \in \{\mathbf{X}^T \boldsymbol{\alpha}, \mathbf{X} \in \Omega_X\}$.

$$\hat{m}_\mu(u) - m_\mu(u) = o_p(n^{-1/4}), \quad \hat{m}_\mu^{(1)}(u) - m_\mu^{(1)}(u) = o_p(n^{-1/4}), \\ \hat{\mathcal{H}}(u) - \mathcal{H}(u) = o_p(n^{-1/4}), \quad \hat{g}(u) - g(u) = o_p(n^{-1/4}), \quad \text{and} \quad \hat{g}^{(1)}(u) - g^{(1)}(u) = o_p(n^{-1/4}).$$

These results imply $\hat{\kappa} - \kappa = o_p(n^{-1/4})$. Thus, Newey's assumption 5.1(ii) holds.

After examining Newey's assumptions 5.1 to 5.3, we apply his Lemma 5.1 and find that $\hat{\omega}_{wls}$ has the same limit distribution as the solution to the equation

$$0 = \sum_{i=1}^n \Psi(\mathcal{H}, m_\mu, \kappa, g; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}_i, \mathbf{X}_i). \quad (\text{A.10})$$

Furthermore, it can be seen that the solution to (A.10) has the same limit distribution as described in the statement of Theorem 3.1. Hence, we complete the proof for asymptotic normality.

Finally, we show the efficiency of $\hat{\omega}$. Let $p_\varepsilon(\varepsilon)$ be the probability density function of ε and $p_\varepsilon^{(1)}(\varepsilon)$ be its first-order derivative with respect to ε . Then, the score function of ω is

$$S_\omega = -\mathbf{J}_\omega \frac{1}{g^2} \left\{ m_\mu^{(1)}(U) \mathbf{X}^T, \mathbf{X}^T \right\}^T p_\varepsilon^{(1)}(\varepsilon) / p_\varepsilon(\varepsilon).$$

For any given function q of $\mathbf{X}$, it can be shown that the nuisance tangent space $\mathcal{P}$, for the three nuisance parameters, $q_X(\mathbf{x})$, $p_\varepsilon(\varepsilon)$, and $m_\mu(U)$, is $\{q(\mathbf{X}) : E(q) = 0, E(\varepsilon q) \text{ is a function of } \mathbf{X} \text{ only}\}$. Furthermore, the orthogonal component of $\mathcal{P}$ is

$$\mathcal{P}^\perp = \left\{ \varepsilon q(\mathbf{X}) / g^2 : E(q/g^2 | U) = 0 \right\}.$$

Subsequently, we apply Bickel *et al.*'s (1993) approach and obtain the following semiparametric efficient score function via equation (9),

$$S_{eff} = \mathbf{J}_\omega \varepsilon \left\{ m_\mu^{(1)}(U) \tilde{\mathbf{X}}^T, \tilde{\mathbf{X}}^T \right\}^T / g^2$$

with $\tilde{\mathbf{X}} = \mathbf{X} - E(\mathbf{X}/g^2 | U)/E(1/g^2 | U)$. It can be seen that $S_{eff} \in \mathcal{P}^\perp$.

For any $\varepsilon q \in \mathcal{P}^\perp$, we have $E(q/g^2 | U) = 0$. Accordingly,

$$E \left\{ (S_\omega - S_{eff})^T \varepsilon q(\mathbf{X}) / g^2 \right\} \\ = \mathbf{J}_\omega E \left(- \left\{ \begin{matrix} m_\mu^{(1)}(U) \mathbf{X} \\ \mathbf{X} \end{matrix} \right\}^T \frac{q}{g^2} E \left\{ \frac{\varepsilon p_\varepsilon^{(1)}(\varepsilon)}{p_\varepsilon(\varepsilon)} \right\} - \frac{1}{g^2} \left[\left\{ \begin{matrix} m_\mu^{(1)}(U) \mathbf{X} \\ \mathbf{X} \end{matrix} \right\}^T q - E \left\{ \begin{matrix} m_\mu^{(1)}(U) E(\mathbf{X}/g^2 | U) / E(1/g^2 | U) \\ E(\mathbf{X}/g^2 | U) / E(1/g^2 | U) \end{matrix} \right\}^T \right] \right)$$

Because $E \left\{ \epsilon p_{\epsilon}^{(1)}(\epsilon) / p_{\epsilon}(\epsilon) \right\} = -1$, it follows that

$$E \left\{ (S_{\omega} - S_{eff})^T \epsilon q(\mathbf{X}) \right\} = \mathbf{J}_{\omega} E \left[\left\{ m_{\mu}^{(1)}(U) E(\mathbf{X}^T / g^2 | U), E(\mathbf{X}^T / g^2 | U) \right\} E(q / g^2 | U) \right] = 0.$$

That is, S_{eff} is the projection of S_{ω} to $\mathcal{D}^{\perp}$, and the estimator $\hat{\omega}$ is therefore efficient (see Bickel *et al.*, 1993), completing the proof.

A.5 Proof of (14) and (16) in Theorem 3.2

The proof is identical to that of (13) and (15), except that $\hat{\omega}_{WLS}$ replaces $\hat{\omega}$ everywhere, and we apply the asymptotic expansion of $n^{1/2}(\hat{\omega}_{WLS} - \omega)$ give in Theorem 3.1. Routine calculations yield Theorem 3.2.

References

- Bickel P. Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *Annals of Statistics*. 1978; 6:266–291.
- Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. The Johns Hopkins University Press; 1993.
- Box G, Hill W. Correcting inhomogeneity of variance with power transformation weighting. *Technometrics*. 1974; 16:385–389.
- Box G, Meyer D. An analysis for unreplicated fractional factorials. *Technometrics*. 1986; 28:11–18.
- Cai TT, Wang L. Adaptive variance function estimation in heteroscedastic nonpara-metric regression. *Annals of Statistics*. 2008; 36:2025–2054.
- Carroll, R.; Ruppert, D. Transformation and Weighting in Regression. Chapman & Hall; New York: 1988.
- Carroll RJ. Adapting for heteroscedasticity in linear models. *Annals of Statistics*. 1982; 10:1224–1233.
- Carroll RJ. Variances are not always nuisance parameters. *Biometrics*. 2003; 59:211–220. [PubMed: 12926705]
- Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association*. 1997; 92:477–489.
- Carroll RJ, Hördle W. Second order effects in semiparametric weighted least squares regression. *Statistics*. 1989; 2:179–186.
- Carroll RJ, Ruppert D. Robust estimation in heteroscedasticity linear models. *Annals of Statistics*. 1982; 10:429–441.
- Carroll, RJ.; Ruppert, D. Transformation and Weighting in Regression. Chapman and Hall; Oxford: 1988.
- Davidian M, Carroll RJ. Variance function estimation. *Journal of the American Statistical Association*. 1987; 82:1079–1091.
- Davidian M, Carroll RJ, Smith W. Variance functions and the minimum detectable concentration in assays. *Biometrika*. 1988; 75:549–556.
- Fuller W, Rao J. Estimation for a linear regression model with unknown diagonal covariance matrix. *Annals of Statistics*. 1978; 6:1149–1158.
- Hall P, Carroll RJ. Variance function estimation in regression - the effect of estimating the mean. *Journal of the Royal Statistical Society Series B-Methodological*. 1989; 51:3–14.
- Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*. 1993; 58:71–120.
- Knight K. Limiting distributions for L_1 regression estimators under general conditions. *Annals of Statistics*. 1998; 26:755–770.

- Liang H, Liu X, Li R, Tsai CL. Estimation and testing for partially linear single-index models. *Annals of Statistics*. 2010; 38:3811–3836. [PubMed: 21625330]
- Ma Y, Chiou J-M, Wang N. Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*. 2006; 93:75–84.
- Ma Y, Zhu L. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *Journal of the Royal Statistical Society, Series B*. 2013a; 75:305–322.
- Ma Y, Zhu L. Efficient estimation in sufficient dimension reduction. *Annals of Statistics*. 2013b; 41:250–268. [PubMed: 24058219]
- Mack YP, Silverman BW. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*. 1982; 61:405–415.
- Newey WK. The asymptotic variance of semiparametric estimators. *Econometrica*. 1994; 62:1349–1382.
- Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions. *Bioinformatics*. 2012; 28:1487–1494. [PubMed: 22492641]
- Thomas, L.; Stefanski, LA.; Davidian, M. Measurement error model methods for bias reduction and variance estimation in logistic regression with estimated variance predictors. Tech. rep. North Carolina State University; 2012.
- Van Keilegom I, Wang L. Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*. 2010; 4:133–160.
- Wang J-L, Xue L, Zhu L, Chong YS. Estimation for a partial-linear single-index model. *Annals of Statistics*. 2010; 38:246–274.
- Western B, Bloome D. Variance function regressions for studying inequality. *Sociological Methodology*. 2009; 39:293–326.
- Xia Y. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*. 2008; 103:1631–1640.
- Xia Y, Tong H, Li WK, Zhu L-X. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B*. 2002; 64:363–410.
- Xia YC, Härdle W. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*. 2006; 97:1162–1184.
- Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*. 2002; 97:1042–1054.

Figure 1.

Estimates for Example 1 of Section 4.2 with $n = 200$. The left column $\hat{m}_\mu$ is the initial mean estimates from Section 2.3. The second left column $\hat{m}_\nu$ is the initial variance estimates from Section 2.4. The second right column $\hat{m}_\mu^e$ are the weighted mean estimates from Section 2.5. The right column, $\hat{m}_\nu^e$ is the updated variance function estimates from Section 2.6. The red solid curves are the true functions. The three rows correspond to $\sigma = 0.2, 0.5$, and 1 respectively.

Figure 2.

Estimates for Example 2 of Section 4.2 with $n = 200$. The left column $\hat{m}_\mu$ is the initial mean estimates from Section 2.3. The second left column $\hat{m}_\nu$ is the initial variance estimates from Section 2.4. The second right column $\hat{m}_\mu^e$ are the weighted mean estimates from Section 2.5. The right column, $\hat{m}_\nu^e$ is the updated variance function estimates from Section 2.6. The red solid curves are the true functions. The three rows correspond to $\sigma = 0.2, 0.5$, and 1 respectively.

Figure 3.

Top-left: the scatterplot of $Y_i - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$ vs. $\mathbf{X}_i^T \hat{\boldsymbol{\alpha}}$ and the solid line shows the fitted $m_{\mu}^{\hat{}}$. Top-right: the scatterplot of $\log(R_i)$ vs. $\hat{m}_{\nu}(\mathbf{X}_i^T \hat{\boldsymbol{\theta}}) + \mathbf{X}_i^T \hat{\boldsymbol{\xi}}$ and the solid line is just a line that goes through the origin with slope one as a reference line. Figures on the second row are similar, but with the second step estimators.

Table 1

Estimates of parameters defined in Sections 2.5-2.6 and their estimated standard errors ($\hat{se}$) for the NMMAPS Example in Section 4.1. Here "Temp" is temperature.

Parameters	Mean Temp	Relative Humidity	Mean CO ₂	Mean PM ₁₀	Mean SO ₂	Humidity Range	Temp Range
$\hat{\alpha}$	0.893	0.240	-0.005	0.320	0.098	-0.149	0.093
$\hat{se}(\hat{\alpha})$	0.004	0.006	0.007	0.012	0.009	0.008	0.005
$\hat{\beta}$	0.322	-0.742	-1.593	-0.330	-0.470	0.353	0.931
$\hat{se}(\hat{\beta})$	0.017	0.027	0.029	0.051	0.038	0.030	0.024
$\hat{\theta}$	0.215	-0.282	0.260	-0.746	-0.048	-0.327	0.372
$\hat{se}(\hat{\theta})$	0.060	0.067	0.065	0.052	0.081	0.051	0.070
$\hat{\zeta}$	0.940	-0.055	-0.748	-0.442	1.065	0.288	-0.558
$\hat{se}(\hat{\zeta})$	0.197	0.250	0.248	0.215	0.297	0.236	0.293

Table 2

RMSE for estimates and the corresponding standard error of MSE (in parentheses) obtained in Example 1 of Section 4.2 for $n = 200$, with $m_\mu(x) = 15 \sin(0.4x)$ and $m_\nu(x) = \cos(0.5x) + 3/2$. We contrast the unweighted and weighted mean estimates in Sections 2.3 and 2.5, respectively, the latter theoretically more efficient. We also contrast the initial and final variance estimates in Sections 2.4 and 2.6, respectively, the latter theoretically more efficient. In addition, we present the infeasible mean estimates that are weighted by the (unknown) true variances, while the infeasible variance estimates are based on residuals from the (unknown) true mean function. The infeasible mean estimates are theoretically asymptotically equivalent to the weighted mean estimate, while the infeasible variance estimates are asymptotically more efficient than the feasible ones.

Unweighted mean estimates, Section 2.3			
σ	α	β	m_μ
0.2	0.052 (0.014)	0.407 (0.108)	0.631 (0.165)
0.5	0.121 (0.041)	0.762 (0.261)	1.279 (0.450)
1	0.250 (0.075)	1.111 (0.506)	2.799 (1.291)

Weighted mean estimates from first stage variance estimates, Section 2.5			
σ	α	β	m_μ
0.2	0.047 (0.014)	0.332 (0.120)	0.627 (0.135)
0.5	0.091 (0.027)	0.575 (0.208)	1.162 (0.281)
1	0.164 (0.054)	0.841 (0.316)	2.192 (0.802)

Infeasible weighted mean estimates from true variances			
σ	α	β	m_μ
0.2	0.045 (0.013)	0.330 (0.083)	0.620 (0.148)
0.5	0.090 (0.026)	0.572 (0.129)	1.163 (0.313)
1	0.162 (0.049)	0.827 (0.217)	2.013 (0.760)

Variance estimates from unweighted mean estimates, Section 2.4			
σ	θ	ζ	m_ν
0.2	0.204 (0.093)	0.171 (0.056)	0.277 (0.092)
0.5	0.191 (0.083)	0.176 (0.055)	0.267 (0.103)
1	0.200 (0.094)	0.174 (0.055)	0.255 (0.074)

Variance estimates from weighted mean estimates, Section 2.6			
σ	θ	ζ	m_ν
0.2	0.199 (0.089)	0.171 (0.055)	0.235 (0.080)
0.5	0.189 (0.094)	0.176 (0.053)	0.253 (0.065)
1	0.189 (0.089)	0.173 (0.054)	0.231 (0.088)

Infeasible variance estimates from true means			
σ	θ	ζ	m_ν
0.2	0.188 (0.085)	0.171 (0.050)	0.187 (0.055)
0.5	0.189 (0.085)	0.172 (0.049)	0.189 (0.055)

Infeasible variance estimates from true means			
σ	θ	ζ	m_v
1	0.188 (0.085)	0.171 (0.050)	0.202 (0.054)

Table 3

RMSE for estimates and the corresponding standard error of MSE (in parentheses) obtained in Example 2 of Section 4.2 for $n = 200$, with $m_\mu(x) = x^2$ and $m_\nu(x) = 2/\{1 + \exp(-2x)\} + 3/2$. Other descriptions are the same as in Table 2.

Unweighted mean estimates, Section 2.3			
σ	α	β	m_μ
0.2	0.100 (0.044)	0.699 (0.149)	1.774 (0.665)
0.5	0.238 (0.079)	0.950 (0.298)	4.508 (1.752)
1	0.330 (0.108)	1.205 (0.524)	8.934 (3.157)

Weighted mean estimates from first stage variance estimates, Section 2.5			
σ	α	β	m_μ
0.2	0.062 (0.027)	0.469 (0.127)	1.629 (0.503)
0.5	0.134 (0.048)	0.775 (0.217)	3.281 (1.259)
1	0.226 (0.077)	1.096 (0.342)	6.443 (3.077)

Infeasible weighted mean estimates from true variances			
σ	α	β	m_μ
0.2	0.052 (0.016)	0.357 (0.098)	1.628 (0.531)
0.5	0.106 (0.042)	0.646 (0.213)	2.992 (1.009)
1	0.187 (0.072)	0.919 (0.309)	5.300 (2.341)

Variance estimates from unweighted mean estimates, Section 2.4			
σ	θ	ζ	m_ν
0.2	0.248 (0.118)	0.238 (0.0817)	0.422 (0.102)
0.5	0.238 (0.125)	0.226 (0.0758)	0.416 (0.114)
1	0.238 (0.131)	0.231 (0.0807)	0.378 (0.109)

Variance estimates from weighted mean estimates, Section 2.6			
σ	θ	ζ	m_ν
0.2	0.235 (0.106)	0.232 (0.067)	0.315 (0.095)
0.5	0.215 (0.094)	0.202 (0.073)	0.274 (0.091)
1	0.198 (0.078)	0.193 (0.058)	0.269 (0.078)

Infeasible variance estimates from true means			
σ	θ	ζ	m_ν
0.2	0.191 (0.065)	0.181 (0.065)	0.201 (0.067)
0.5	0.192 (0.065)	0.181 (0.065)	0.199 (0.066)
1	0.191 (0.064)	0.182 (0.062)	0.200 (0.068)

Table 4

Standard errors for the second-stage estimators in Example 1 of Section 4.2, where for each parameter the numbers on the first line are the estimated standard errors and the numbers on the second line are the Monte Carlo standard errors.

$\sigma = 0.2$								
$\hat{\text{se}}(\hat{\alpha})$	0.017	0.013	0.017	0.013	0.015	0.017	0.019	0.016
	0.018	0.015	0.015	0.015	0.019	0.018	0.017	0.014
$\hat{\text{se}}(\hat{\beta})$	0.135	0.125	0.122	0.124	0.102	0.124	0.138	0.134
	0.134	0.103	0.126	0.102	0.118	0.118	0.126	0.114
$\hat{\text{se}}(\hat{\theta})$	0.075	0.065	0.073	0.066	0.070	0.076	0.077	0.080
	0.066	0.064	0.095	0.075	0.060	0.074	0.075	0.081
$\hat{\text{se}}(\hat{\zeta})$	0.053	0.050	0.056	0.049	0.058	0.052	0.053	0.054
	0.073	0.060	0.064	0.059	0.058	0.054	0.055	0.065
$\sigma = 0.5$								
$\hat{\text{se}}(\hat{\alpha})$	0.027	0.029	0.029	0.029	0.025	0.027	0.025	0.025
	0.035	0.032	0.033	0.031	0.036	0.032	0.033	0.029
$\hat{\text{se}}(\hat{\beta})$	0.282	0.209	0.202	0.202	0.200	0.151	0.147	0.154
	0.299	0.171	0.230	0.165	0.200	0.183	0.183	0.182
$\hat{\text{se}}(\hat{\theta})$	0.077	0.073	0.074	0.067	0.071	0.076	0.078	0.077
	0.055	0.061	0.093	0.081	0.060	0.060	0.068	0.075
$\hat{\text{se}}(\hat{\zeta})$	0.060	0.053	0.049	0.050	0.058	0.053	0.053	0.054
	0.079	0.062	0.065	0.063	0.059	0.056	0.053	0.069
$\sigma = 1$								
$\hat{\text{se}}(\hat{\alpha})$	0.036	0.050	0.041	0.047	0.045	0.063	0.055	0.050
	0.066	0.062	0.055	0.059	0.063	0.060	0.055	0.048
$\hat{\text{se}}(\hat{\beta})$	0.287	0.280	0.246	0.271	0.303	0.238	0.275	0.291
	0.373	0.248	0.262	0.197	0.268	0.212	0.226	0.241
$\hat{\text{se}}(\hat{\theta})$	0.078	0.067	0.070	0.066	0.081	0.075	0.078	0.079
	0.060	0.060	0.075	0.081	0.060	0.063	0.069	0.069
$\hat{\text{se}}(\hat{\zeta})$	0.052	0.048	0.047	0.046	0.057	0.059	0.052	0.051
	0.065	0.061	0.061	0.062	0.061	0.060	0.057	0.074

Table 5

Standard errors for the second-stage estimators in Example 2 of Section 4.2, where for each parameter the numbers on the first line are the estimated standard errors and the numbers on the second line are the Monte Carlo standard errors.

$\sigma = 0.2$								
$\hat{\text{se}}(\hat{\alpha})$	0.020	0.020	0.021	0.021	0.021	0.021	0.022	0.022
	0.023	0.023	0.026	0.027	0.022	0.019	0.021	0.025
$\hat{\text{se}}(\hat{\beta})$	0.177	0.166	0.173	0.148	0.163	0.165	0.181	0.169
	0.245	0.136	0.149	0.137	0.171	0.152	0.184	0.154
$\hat{\text{se}}(\hat{\theta})$	0.083	0.071	0.080	0.080	0.088	0.077	0.086	0.081
	0.099	0.076	0.078	0.091	0.093	0.099	0.084	0.069
$\hat{\text{se}}(\hat{\zeta})$	0.081	0.062	0.072	0.075	0.081	0.062	0.055	0.077
	0.095	0.062	0.070	0.066	0.074	0.070	0.057	0.075
$\sigma = 0.5$								
$\hat{\text{se}}(\hat{\alpha})$	0.048	0.044	0.046	0.048	0.043	0.043	0.019	0.041
	0.052	0.047	0.038	0.055	0.057	0.049	0.053	0.045
$\hat{\text{se}}(\hat{\beta})$	0.370	0.292	0.250	0.247	0.260	0.278	0.243	0.318
	0.415	0.217	0.196	0.202	0.256	0.256	0.257	0.223
$\hat{\text{se}}(\hat{\theta})$	0.084	0.076	0.070	0.078	0.087	0.072	0.079	0.074
	0.081	0.091	0.093	0.088	0.076	0.079	0.065	0.070
$\hat{\text{se}}(\hat{\zeta})$	0.078	0.065	0.066	0.075	0.077	0.062	0.078	0.070
	0.086	0.062	0.079	0.063	0.071	0.065	0.060	0.070
$\sigma = 1$								
$\hat{\text{se}}(\hat{\alpha})$	0.088	0.060	0.060	0.066	0.063	0.060	0.062	0.069
	0.108	0.084	0.065	0.079	0.087	0.079	0.073	0.083
$\hat{\text{se}}(\hat{\beta})$	0.380	0.281	0.316	0.291	0.290	0.277	0.300	0.386
	0.517	0.257	0.236	0.243	0.259	0.275	0.268	0.246
$\hat{\text{se}}(\hat{\theta})$	0.091	0.079	0.076	0.082	0.091	0.076	0.083	0.083
	0.075	0.074	0.080	0.068	0.069	0.077	0.071	0.080
$\hat{\text{se}}(\hat{\zeta})$	0.079	0.077	0.068	0.079	0.075	0.065	0.079	0.078
	0.089	0.054	0.071	0.061	0.068	0.062	0.058	0.069