ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2015)
77, *Part* 1, *pp.* 171–194

Variance function partially linear single-index models

Heng Lian,

Nanyang Technological University, Singapore

Hua Liang

George Washington University, Washington DC, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

[Received December 2012. Final revision December 2013]

Summary. We consider heteroscedastic regression models where the mean function is a partially linear single-index model and the variance function depends on a generalized partially linear single-index model. We do not insist that the variance function depends only on 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 non-parametric parts of the model is developed. Simulations illustrate the results. An empirical example involving ozone levels is used to illustrate the results further and is shown to be a case where the variance function does not depend on the mean function.

Keywords: Asymptotic theory; Estimating equation; Identifiability; Kernel regression; Modelling 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 on a generalized partially linear single-index model. We do not insist that the variance function depends only on the mean function, as happens in the classical generalized partially linear single-index model. Our model is

$$Y = \mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu) + g\{v(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\zeta}, m_v)\}\epsilon, \tag{1}$$

$$\mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu) = m_\mu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}) + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2}$$

$$v(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\zeta}, m_v) = m_v(\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}) + \mathbf{X}^{\mathrm{T}}\boldsymbol{\zeta}, \tag{3}$$

where $g(\cdot)$ is a known function, whereas $m_\mu(\cdot)$ and $m_v(\cdot)$ are two unknown smooth functions, ϵ is independent of $\mathbf{X}$, $E(\epsilon) = 0$ and $E(|\epsilon|) = 1$, the last condition being 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 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta} = \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\zeta} = 0$.

Address for correspondence: Hua Liang, Department of Statistics, George Washington University, Washington DC 20052, USA.
E-mail: hliang@gwu.edu

1369–7412/15/77171

This model retains the flexibility of a non-parametric regression model but has dimension reduction ability to avoid fitting a multivariate non-parametric regression function. Methods for estimating the mean function $\mu(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)$ in model (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 (Xia *et al*., 2002). The partially linear structure that is specified in equation (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, whereas Liang *et al*. (2010) proposed a profile least squares estimation procedure. Although they allowed heteroscedasticity, they did not explicitly model it; instead they simply assumed that $E(Y|\mathbf{X}, \mathbf{Z}) = m_\mu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}) + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}$, and their asymptotic theory assumed that $v(\mathbf{X}, \mathbf{Z}) \equiv 1$, although this improvement changes only their asymptotic covariance matrix. Moreover, all existing work mainly focuses on estimation of the mean parameters $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$.

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 improve 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) and 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 who had recently been 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 modelling heteroscedasticity in linear or non-linear 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 be a parametric function of the covariates, and a non-parametric approach (Carroll and Härdle, 1989; Fuller and Rao, 1978; Hall and Carroll, 1989) with a fully non-parametric 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, which is 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 that they used 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 $g_*\{m_*(\mathbf{X}^{\mathrm{T}}\mathbf{a}) + \mathbf{X}^{\mathrm{T}}\mathbf{b}\}$ with $\|\mathbf{a}\| = 1$ and $\mathbf{a}^{\mathrm{T}}\mathbf{b} = 0$ 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 that had been 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 that was 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. Although the techniques in Thomas *et al.* (2012) and Teschendorff and Widschwendter (2012) 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 of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, and then the variance function $g\{v(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\zeta},m_v)\}$. As a by-product, we can do reweighting (generalized least squares) to improve the estimates of $(\boldsymbol{\alpha},\boldsymbol{\beta})$. It is worth mentioning that the model that $\boldsymbol{\alpha}=\boldsymbol{\theta},\boldsymbol{\beta}=\boldsymbol{\zeta}$ 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 generalized linear models. The model (1)–(3) is also interesting conceptually because it suggests that, if either $\boldsymbol{\alpha}\neq\boldsymbol{\theta}$ or $\boldsymbol{\beta}\neq\boldsymbol{\zeta}$, then different linear combinations of $\mathbf{X}$ are governing the distribution of Y. This can be important in classification, for example, because then extra information is available when modelling 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 on the mean function. All technical assumptions and proofs of the main results are in Appendix A.

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 shall 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:

(a) initial estimation of the mean function ignoring the heteroscedasticity;
(b) using the absolute residuals from the initial estimate of the mean function, develop an initial estimate of the variance function;
(c) 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;
(d) 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 $\|\boldsymbol{\alpha}\|=1$ (with first non-zero component positive) and $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta}=0$ in effect reduce the number of parameters $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ 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 that the first component of $\boldsymbol{\alpha}$, α_1, is positive, and thus we can write $\boldsymbol{\alpha} = ((1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{1/2}, \alpha_2, \ldots, \alpha_p)^{\mathrm{T}}$ where $\boldsymbol{\alpha}_{\backslash 1} = (\alpha_2, \ldots, \alpha_p)^{\mathrm{T}}$ is $\boldsymbol{\alpha}$ without the first component. Similarly we denote $\boldsymbol{\beta}_{\backslash 1} = (\beta_2, \ldots, \beta_p)^{\mathrm{T}}$. Because of the second constraint, we have

$$\beta_1 = -\alpha_1^{-1}\left(\sum_{j=2}^{p} \alpha_j \beta_j\right) = -(1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{-1/2}\left(\sum_{j=2}^{p} \alpha_j \beta_j\right).$$

Thus $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is a function of $\omega := (\alpha_2, \ldots, \alpha_p, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$. The $2p \times (2p-2)$ Jacobian matrix $\partial(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}} / \partial\omega$ is

$$\mathbf{J}_\omega = \begin{pmatrix} -\dfrac{\boldsymbol{\alpha}_{\backslash 1}^{\mathrm{T}}}{(1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{1/2}} & \mathbf{0}_{1 \times (p-1)} \\ I_{(p-1)\times(p-1)} & \mathbf{0}_{(p-1)\times(p-1)} \\ -\dfrac{\boldsymbol{\beta}_{\backslash 1}^{\mathrm{T}}}{(1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{1/2}} - (\boldsymbol{\alpha}_{\backslash 1}^{\mathrm{T}} \boldsymbol{\beta}_{\backslash 1}) \dfrac{\boldsymbol{\alpha}_{\backslash 1}^{\mathrm{T}}}{(1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{3/2}} & -\dfrac{\boldsymbol{\alpha}_{\backslash 1}^{\mathrm{T}}}{(1 - \|\boldsymbol{\alpha}_{\backslash 1}\|^2)^{1/2}} \\ \mathbf{0}_{(p-1)\times(p-1)} & I_{(p-1)\times(p-1)} \end{pmatrix},$$

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

In the same way, we assume that the first component of $\boldsymbol{\theta}$, θ_1, is positive, and thus we can introduce $\boldsymbol{\theta}_{\backslash 1}$ and $\boldsymbol{\zeta}_{\backslash 1}$. Write $\vartheta = (\theta_2, \ldots, \theta_p, \zeta_2, \ldots, \zeta_p)^{\mathrm{T}}$, and define $\mathbf{J}_\vartheta$ similarly to $\mathbf{J}_\omega$.

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, be independent samples of $(Y, \mathbf{X})$. In what follows, we set $R_i = |Y_i - \mu(\mathbf{X}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, m_\mu)|$, $\varepsilon_i = Y_i - \{m_\mu(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}) + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}$, $d_i = I_{(\varepsilon_i > 0)} - I_{(\varepsilon_i \leqslant 0)} = \mathrm{sgn}(\varepsilon_i)$, $g_i = g\{m_v(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}) + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\zeta}\}$, $\delta_i = R_i - g_i$, $U_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}$ and $T_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}$. Then $E(|\varepsilon_i\|\mathbf{X}_i) = E(R_i|\mathbf{X}_i) = g\{m_v(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}) + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\zeta}\}$. Let $g^{(1)}$ denote the first derivative of g; define $g^{(1)2} = \{g^{(1)}\}^2$. Set $\Lambda_\mu = (m_\mu^{(1)}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha})\mathbf{X}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $\Lambda_{i\mu} = (m_\mu^{(1)}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha})\mathbf{X}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}})$, $\Lambda_v = (m_v^{(1)}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta})\mathbf{X}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $\Lambda_{iv} = (m_v^{(1)}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})\mathbf{X}_i^{\mathrm{T}}, \mathbf{X}_i^{\mathrm{T}})$, $\tilde{\Lambda}_\mu = \Lambda_\mu - E(\Lambda_\mu/g^2|\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha})/E(1/g^2|\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha})$, $\tilde{\Lambda}_{i\mu} = \Lambda_{i\mu} - E(\Lambda_{i\mu}/g_i^2|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha})/E(1/g_i^2|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha})$, $\tilde{\Lambda}_v = \Lambda_v - E\{\Lambda_v g^{(1)2}(\cdot)|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}\}/E\{g^{(1)2}(\cdot)|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}\}$ and $\tilde{\Lambda}_{iv} = \Lambda_{iv} - E\{\Lambda_{iv} g_i^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}/E\{g_i^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}$, and $A^{\otimes 2} = AA^{\mathrm{T}}$ for any matrix A. We shall estimate the two unknown functions $m_\mu(u)$ and $m_v(t)$ non-parametrically. For notational simplicity, we use the same kernel function and bandwidth for both non-parametric 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 model (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 $Y - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$ on $\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}$, resulting in $\hat{m}_\mu(\cdot)$. More precisely, it should be $\hat{m}_\mu(\cdot, \boldsymbol{\alpha}, \boldsymbol{\beta})$, but to keep the notation simpler we shall sometimes use the convention $\hat{m}_\mu(\cdot)$. Then $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are estimated by profiling. Thus, for any given $(\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, we use local linear regression to estimate m_μ by minimizing

$$\sum_{i=1}^{n} [Y_i - \{a + b(X_i^{\mathrm{T}}\boldsymbol{\alpha} - u) + X_i^{\mathrm{T}}\boldsymbol{\beta}\}]^2 K_h(X_i^{\mathrm{T}}\alpha - u) \tag{4}$$

with respect to a and b, where 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 expression (4) and set $\hat{m}_\mu(u) = \hat{a}$.

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

$$\sum_{i=1}^{n} [Y_i - \{\hat{m}_\mu(X_i^{\mathrm{T}}\boldsymbol{\alpha}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + X_i^{\mathrm{T}}\boldsymbol{\beta}\}]^2, \tag{5}$$

with respect to $\boldsymbol{\alpha}_{\backslash 1}$ and $\boldsymbol{\beta}_{\backslash 1}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are functions of $\boldsymbol{\alpha}_{\backslash 1}$ and $\boldsymbol{\beta}_{\backslash 1}$. We define the solution of problem (5) as $\hat{\omega}=(\hat{\boldsymbol{\alpha}}_{\backslash 1}^{\mathrm{T}},\hat{\boldsymbol{\beta}}_{\backslash 1}^{\mathrm{T}})^{\mathrm{T}}$. 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 we use $\mathbf{J}_{\omega}$ for $\mathbf{J}_{\omega}$ evaluated at the true value $(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)=\{\alpha(\omega_0),\beta(\omega_0)\}$.

More efficiency for estimating $(m_{\mu},\boldsymbol{\alpha},\boldsymbol{\beta})$ can be obtained via generalized least squares, which we shall 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 distinguished 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 they called this potential efficiency gain 'tenuous' because it is less robust to outliers. Here we use absolute residuals and follow a profiling approach that is analogous to that in Section 2.3. Of course, if one chooses to use squared residuals, the same algorithm as described below applies, by replacing $\hat{R}_i$ by $\hat{R}_i^2$ and $g(\cdot)$ by $g^2(\cdot)$ in equations (6) and (7) below. The asymptotic distribution results for $(\boldsymbol{\theta},\boldsymbol{\zeta})$ that are described in theorem 2 below are also easily modified.

Define $R=|Y-\mu(\mathbf{X},\boldsymbol{\alpha},\boldsymbol{\beta},m_{\mu})|$ and $\hat{R}=|Y-\mu(\mathbf{X},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{m}_{\mu})|$. Recall that $E(|\epsilon|)=1$. Then, approximately, $E(\hat{R}|\mathbf{X})\approx g\{v(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\zeta},m_v)\}$. Thus, write $\hat{R}_i=|Y_i-\mu(\mathbf{X}_i,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}},\hat{m}_{\mu})|=|Y_i-\{\hat{m}_{\mu}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\alpha}})+\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\}|$ and $\hat{S}_i=\{\hat{m}_{\mu}(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\alpha}})+\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\}-\{m_{\mu}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha})+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}$.

We first minimize in (a_0,a_1)

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

resulting in $\hat{m}_v(\cdot,\boldsymbol{\theta},\boldsymbol{\zeta})$. We then estimate $(\boldsymbol{\theta},\boldsymbol{\zeta})$ by minimizing

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

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

2.5. More efficient estimation of the mean function

We now investigate how we can more efficiently estimate $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 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 equation (4) by

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

We then update the estimates of $(\boldsymbol{\alpha},\boldsymbol{\beta})$ by the same process, but the least squares minimization is 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 $(\boldsymbol{\alpha},\boldsymbol{\beta})$ 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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by minimizing

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

with respect to $\boldsymbol{\alpha}_{\backslash 1}$ and $\boldsymbol{\beta}_{\backslash 1}$. We define the solution of problem (9) as $\hat{\omega}_{\mathrm{WLS}}=(\hat{\boldsymbol{\alpha}}_{w,\backslash 1}^{\mathrm{T}},\hat{\boldsymbol{\beta}}_{w,\backslash 1}^{\mathrm{T}})^{\mathrm{T}}$. Thus the weighted generalized least squared estimator for $(\boldsymbol{\alpha},\boldsymbol{\beta})$ is $(\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}},\hat{\boldsymbol{\beta}}_{\mathrm{WLS}})=\{\boldsymbol{\alpha}(\hat{\omega}_{\mathrm{WLS}}),\boldsymbol{\beta}(\hat{\omega}_{\mathrm{WLS}})\}$.

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 $\hat{R}=|Y-\mu\{\mathbf{X},\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}},\hat{\boldsymbol{\beta}}_{\mathrm{WLS}},\hat{m}(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}},\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}},\hat{\boldsymbol{\beta}}_{\mathrm{WLS}})\}|$. We denote the resulting estimators as $\hat{\vartheta}_{\mathrm{WLS}}$ and $\hat{m}_{v,\mathrm{WLS}}$.

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 $g_*\{m_*(\mathbf{X}^{\mathrm{T}}\mathbf{a})+\mathbf{X}^{\mathrm{T}}\mathbf{b}\}$ with $\|\mathbf{a}\|=1$ and $\mathbf{a}^{\mathrm{T}}\mathbf{b}=0$, 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 var(ϵ)

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

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

Under the conditions of theorem 1 in Section 3.1 and assuming that $E(\epsilon^4)<\infty$, a straightforward but tedious manipulation shows that $n^{1/2}(\hat{\sigma}_n^2-\sigma^2)\rightarrow N[0,E\{(\epsilon^2-\sigma^2)^2\}]$.

3. Asymptotic results

3.1. Mean function estimation

Theorem 1. Suppose that assumptions 1 and 2 in Appendix A hold. Define $Q_\omega=E\{(\mathbf{J}_\omega^{\mathrm{T}}\tilde{\Lambda}_\mu)^{\otimes 2}\}$. As $n\rightarrow\infty$, $nh^4\rightarrow\infty$ and $nh^6\rightarrow 0$, for the unweighted mean function estimates of Section 2.3,

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

Accordingly, $n^{1/2}((\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})^{\mathrm{T}},(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathrm{T}})^{\mathrm{T}}\rightarrow N[0,\,E\{\varepsilon^2\mathbf{J}_\omega Q_\omega^{-1}E(g\mathbf{J}_\omega^{\mathrm{T}}\tilde{\Lambda}_\mu)^{\otimes 2}Q_\omega^{-1}\mathbf{J}_\omega^{\mathrm{T}}\}]$. Further, for the weighted mean function estimates of Section 2.5, Define $Q_{w,\omega}=E\{(\mathbf{J}_\omega^{\mathrm{T}}\tilde{\Lambda}_\mu/g)^{\otimes 2}\}$. Then

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

$$n^{1/2} Q_{w,\omega}(\hat{\omega}_{\mathrm{WLS}} - \omega) = n^{-1/2} \sum_{i=1}^{n} \frac{\epsilon_i \mathbf{J}_{\omega}^{\mathrm{T}} \tilde{\Lambda}_{i\mu}}{g_i^2} + o_p(1). \tag{10}$$

Accordingly, $n^{1/2}\{(\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}} - \boldsymbol{\alpha})^{\mathrm{T}}, (\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} - \boldsymbol{\beta})^{\mathrm{T}}\}^{\mathrm{T}} \rightarrow N\{0, E(\epsilon^2 \mathbf{J}_{\omega} Q_{w,\omega}^{-1} \mathbf{J}_{\omega}^{\mathrm{T}})\}$. 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 three implications of theorem 1. First and most obvious, the weighted estimates $\hat{\omega}_{\mathrm{WLS}}$ are more efficient than their unweighted version $\hat{\omega}$. Second, the function estimated, $\hat{m}_{\mu}(u, \hat{\boldsymbol{\alpha}}_{\mathrm{WLS}})$, is also more efficient than its unweighted version $\hat{m}_{\mu}(u, \hat{\boldsymbol{\alpha}})$. Finally, an implication of theorem 1 is that $\hat{\omega}_{\mathrm{WLS}}$ 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 non-parametric 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 $(\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}}, \hat{\boldsymbol{\beta}}_{\mathrm{WLS}})$ that is described in theorem 1 has an efficiency property even when ϵ is not normally distributed. The asymptotic distribution that is described there has a seemingly different expression from the asymptotic variance from the estimators that were derived by Ma and Zhu (2013a), which were 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 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 2. Suppose that assumptions 1 and 2 in Appendix A hold. Define

$$Q_{\vartheta} = E(\mathbf{J}_{\vartheta}^{\mathrm{T}} \tilde{\Lambda}_{v} g^{(1)})^{\otimes 2}.$$

Define

$$\Sigma_{\vartheta} = E[\varepsilon\{\mathbf{J}_{\vartheta}^{\mathrm{T}} E(g^{(1)} \Lambda_v d | U) + (\mathbf{J}_{\vartheta} - \mathbf{J}_{\omega})^{\mathrm{T}} E(g^{(1)} \Lambda_v d \tilde{\Lambda}_{\mu}) \mathbf{J}_{\omega} Q_{\omega}^{-1} \mathbf{J}_{\omega}^{\mathrm{T}} \tilde{\Lambda}_{\mu}\} + \delta \mathbf{J}_{\vartheta}^{\mathrm{T}} \tilde{\Lambda}_{v} g^{(1)}]^{\otimes 2}, \tag{11}$$

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

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,

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

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

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

Finally, for the parameter estimators,

$$n^{1/2}\,Q_\vartheta(\hat{\vartheta}-\vartheta)\rightarrow N(0,\Sigma_\vartheta),\qquad(15)$$

$$n^{1/2}\,Q_\vartheta(\hat{\vartheta}_{\mathrm{WLS}}-\vartheta)\rightarrow N(0,\Sigma_{\vartheta,\mathrm{WLS}}).\qquad(16)$$

Remark 3. Recalling that $d=\mathrm{sgn}(\epsilon)$, 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, $\Sigma_\vartheta=\Sigma_{\vartheta,\mathrm{WLS}}$. In addition, the last terms in equations (13) and (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 $(\hat{\boldsymbol{\alpha}}_{\mathrm{WLS}},\hat{\boldsymbol{\beta}}_{\mathrm{WLS}},\hat{m}_{\mathrm{WLS}})$ 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 'National morbidity and mortality air pollution study' database, which contains daily mortality, weather and pollution data for 1987–2000. Here we consider data for only the year 1997. We shall use the partially linear single-index model to explore the relationship between daily mean ozone level and some predictors. The selected seven explanatory variables include mean temperature, relative humidity, mean carbon dioxide (CO_2) level, mean PM_{10}-level, mean sulphur dioxide (SO_2) level, daily humidity range and daily temperature range. After excluding 1 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 to those in section 7 of Carroll *et al.* (1997).

Table 1. Estimates of parameters defined in Sections 2.5 and 2.6 and their estimated standard errors $\widehat{\mathrm{se}}$ for the 'National morbidity and mortality air pollution study' example in Section 4.1

Parameter	*Mean temperature*	*Relative humidity*	*Mean CO_2*	*Mean PM_{10}*	*Mean SO_2*	*Humidity range*	*Temperature range*
$\hat{\boldsymbol{\alpha}}$	0.893	0.240	−0.005	0.320	0.098	−0.149	0.093
$\widehat{\mathrm{se}}(\hat{\boldsymbol{\alpha}})$	0.004	0.006	0.007	0.012	0.009	0.008	0.005
$\hat{\boldsymbol{\beta}}$	0.322	−0.742	−1.593	−0.330	−0.470	0.353	0.931
$\widehat{\mathrm{se}}(\hat{\boldsymbol{\beta}})$	0.017	0.027	0.029	0.051	0.038	0.030	0.024
$\hat{\boldsymbol{\theta}}$	0.215	−0.282	0.260	−0.746	−0.048	−0.327	0.372
$\widehat{\mathrm{se}}(\hat{\boldsymbol{\theta}})$	0.060	0.067	0.065	0.052	0.081	0.051	0.070
$\hat{\boldsymbol{\zeta}}$	0.940	−0.055	−0.748	−0.442	1.065	0.288	−0.558
$\widehat{\mathrm{se}}(\hat{\boldsymbol{\zeta}})$	0.197	0.250	0.248	0.215	0.297	0.236	0.293

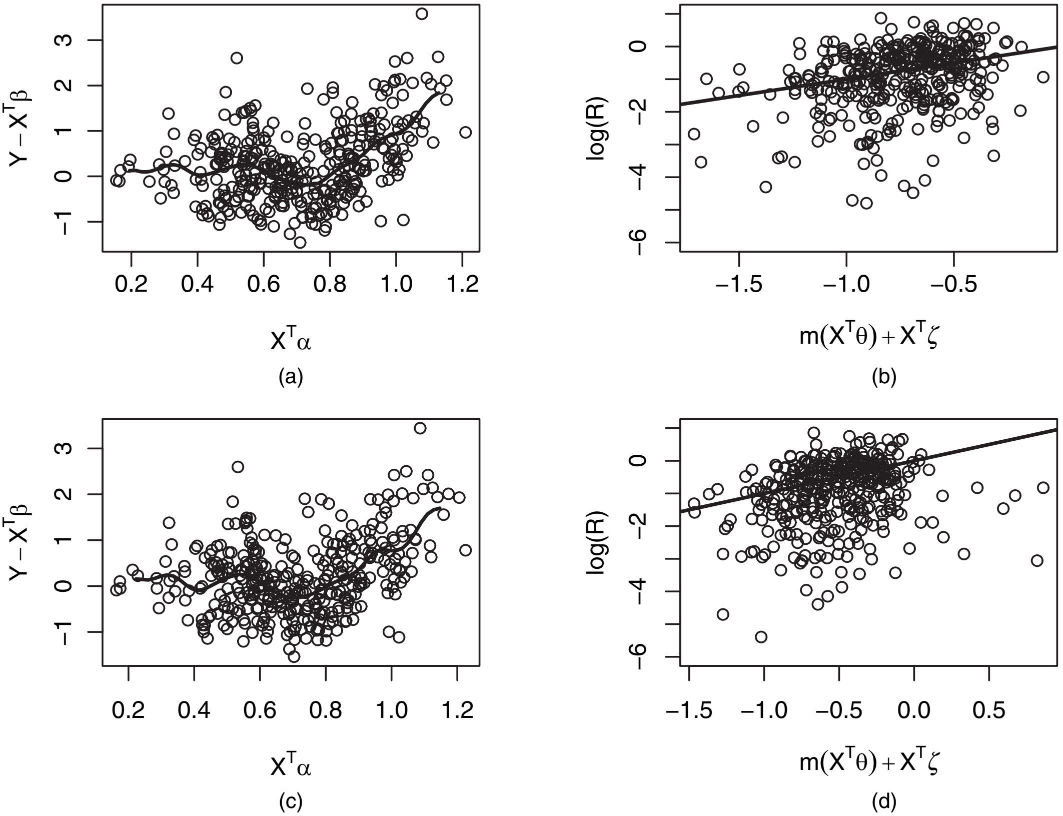

Fig. 1. (a) Scatter plot of $Y_i - \mathbf{X}_i^T\hat{\beta}$ *versus* $\mathbf{X}_i^T\hat{\alpha}$ (———, fitted $\hat{m}_\mu$); (b) scatter plot of $\log(\hat{R}_i)$ *versus* $m_v(\mathbf{X}_i^T\zeta) + \mathbf{X}_i^T\zeta$ (/, reference line with slope 1 through the origin); (c) similar to (a) but with the second-step estimators; (d) similar to (b) but with the second-step estimators

There are some striking conclusions from the analysis. First, every parameter that is associated with the mean function is highly statistically significantly different from 0. In addition, all the single components of the single-index variance parameter $\boldsymbol{\theta}$, except mean SO_2 level, are highly statistically significant. For the partially linear parameter $\boldsymbol{\zeta}$, the coefficients that are 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 on inspection, or by a formal hypothesis test, that $\boldsymbol{\alpha} \neq \boldsymbol{\theta}$ and that $\boldsymbol{\beta} \neq \boldsymbol{\zeta}$. Thus, the variance of Y given $\mathbf{X}$ is not a function of the mean.

Now consider the mean function. In Figs 1(a) and 1(c), we show the scatter plot of $Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}$ *versus* $\mathbf{X}_i^T\hat{\boldsymbol{\alpha}}$ for the first- and second-stage estimators respectively. In both cases, the full curve is the fitted link $\hat{m}_\mu(\cdot)$ for the mean function. It is evident that the means are not constant. There is a hint that $\hat{m}_\mu(\cdot)$ might be constant for smaller values of $\mathbf{X}^T\boldsymbol{\alpha}$, 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 we would expect that $\log(\hat{R})$ would be roughly linear when plotted against $\hat{m}_v(\cdot) + \mathbf{X}^T\hat{\boldsymbol{\zeta}}$. In Figs 1(b) and 1(d), we show the scatter plot of $\log(\hat{R}_i)$ *versus* $\hat{m}_v(\mathbf{X}_i^T\hat{\boldsymbol{\theta}}) + \mathbf{X}_i^T\hat{\boldsymbol{\zeta}}$ for the first- and second-stage estimators respectively, where the full line is just the line that goes through the origin with slope 1

Table 2. RMSE for estimates and the corresponding standard error of the mean-squared error (in parentheses) obtained in example 1 of Section 4.2 for $n = 200$, with $m_\mu(x) = 15\sin(0.4x)$ and $m_v(x) = \cos(0.5x) + \frac{3}{2}$†

σ	α	β	m_μ
Unweighted mean estimates, Section 2.3			
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			
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			
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)
	θ	ζ	m_v
Variance estimates from unweighted mean estimates, Section 2.4			
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			
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			
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)
1	0.188 (0.085)	0.171 (0.050)	0.202 (0.054)

†We contrast the unweighted and weighted mean estimates in Sections 2.3 and 2.5 respectively, the latter being theoretically more efficient. We also contrast the initial and final variance estimates in Sections 2.4 and 2.6 respectively, the latter being theoretically more efficient. In addition, we present the infeasible mean estimates that are weighted by the (unknown) true variances, whereas 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 estimates, whereas the infeasible variance estimates are asymptotically more efficient than the feasible estimates.

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 model (1)–(3), with $g(x) = \exp(x)$ and $\epsilon \sim N(0, \sigma^2)$. The covariates $\mathbf{X} = (X_1, \ldots, X_8)^{\mathrm{T}}$ are generated from a multivariate Gaussian distribution with covariance

Table 3. RMSE for estimates and the corresponding standard error of the mean-squared error (in parentheses) obtained in example 2 of Section 4.2 for $n = 200$, with $m_V(x) = x^2$ and $m_V(x) = 2/\{1 + \exp(-2x)\} + \frac{3}{2}$†

σ	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	m_μ
Unweighted mean estimates, Section 2.3			
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			
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			
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)
	$\boldsymbol{\theta}$	$\boldsymbol{\zeta}$	m_v
Variance estimates from unweighted mean estimates, Section 2.4			
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			
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			
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)

†Other descriptions are the same as in Table 2.

given by $\text{cov}(X_i, X_j) = 0.2^{|i-j|}$. We set $\boldsymbol{\alpha} = (1.0, 1.0, 1.0, 1.0, 0.5, 0.5, 0.5, 0.5)^{\text{T}}$, $\boldsymbol{\beta} = (1.0, -1.0, 1.0, -1.0, 1.0, -1.0, 1.0, -1.0)^{\text{T}}$, $\boldsymbol{\theta} = (0.5, -1, 0.5, -1, -0.5, -1, -1, -0.5)^{\text{T}}$ and $\boldsymbol{\zeta} = (0.2, 0, -0.2, 0, 0.2, 0, -0.25, 0)^{\text{T}}$. As presented above, $\boldsymbol{\alpha}^{\text{T}}\boldsymbol{\beta} = \boldsymbol{\theta}^{\text{T}}\boldsymbol{\zeta} = 0$ but $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are not normalized to have unit norm, although $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are normalized for estimation.

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

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

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 that were described in the outline in Section 2.1 and defined in Sections 2.3–2.6, along with two infeasible estimators:

(a) the unweighted least squares mean function estimates, Section 2.3;
(b) the initial variance function estimates starting from the unweighted least squares mean function estimates, Section 2.4;

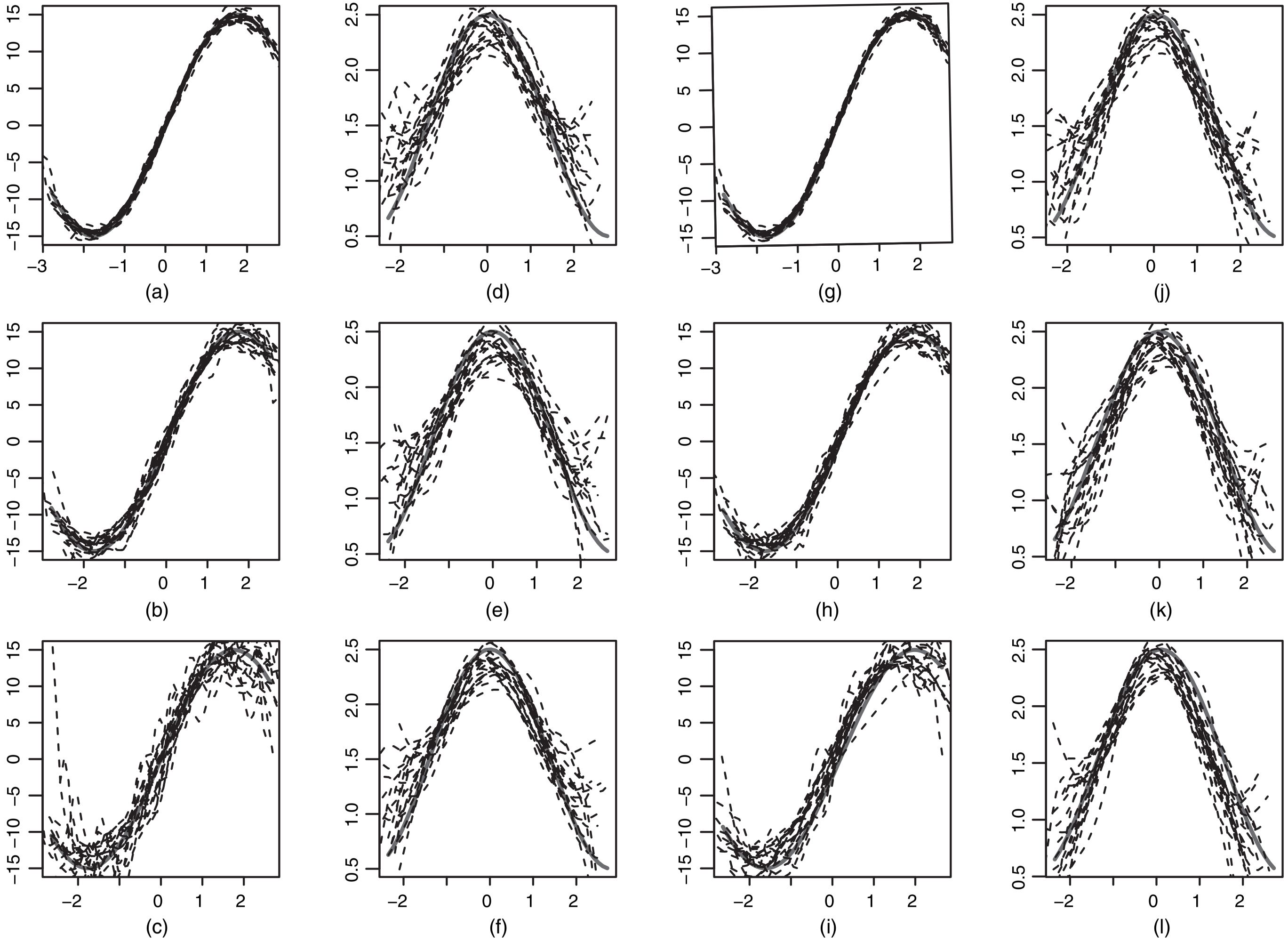

Fig. 2. Estimates for example 1 of Section 4.2 with $n = 200$ (——, true functions): (a)–(c) $\hat{m}_\mu$, initial mean estimate from Section 2.4; (d)–(f) $\hat{m}_V$, initial variance estimate from Section 2.4; (g)–(i) $\hat{m}^{\mathrm{e}}_\mu$, weighted mean estimate from Section 2.5; (j)–(l) $\hat{m}^{\mathrm{e}}_V$, the updated variance function estimates from Section 2.6; (a), (d), (g), (j) $\sigma = 0.2$; (b), (e), (h), (k) $\sigma = 0.5$; (c), (f), (i), (l) $\sigma = 1$

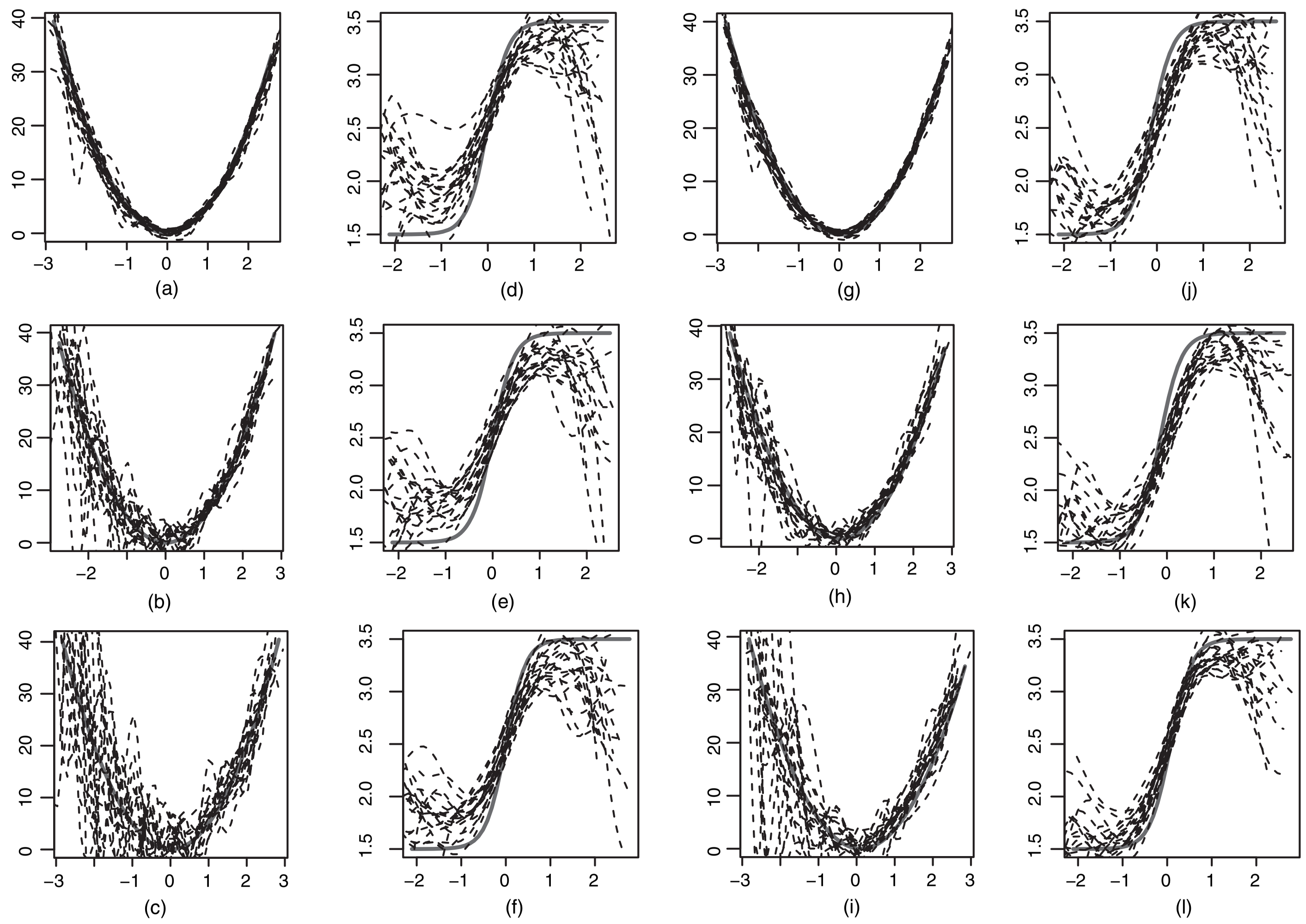

Fig. 3. Estimates for example 2 of Section 4.2 with $n = 200$ (——, true functions): (a)–(c) $\hat{m}_\mu$, initial mean estimate from Section 2.4; (d)–(f) $\hat{m}_V$, initial variance estimate from Section 2.4; (g)–(i) $\hat{m}_\mu^{\mathrm{e}}$, weighted mean estimate from Section 2.5; (j)–(l) $\hat{m}_V^{\mathrm{e}}$, updated variance function estimate from Section 2.6; (a), (d), (g), (j) $\sigma = 0.2$; (b), (e), (h), (k) $\sigma = 0.5$; (c), (f), (i), (l) $\sigma = 1$

Table 4. Standard errors for the second-stage estimators in example 1 of Section 4.2†

$\sigma = 0.2$								
$\widehat{\mathrm{se}}(\hat{\alpha})$	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
$\widehat{\mathrm{se}}(\hat{\beta})$	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
$\widehat{\mathrm{se}}(\hat{\theta})$	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
$\widehat{\mathrm{se}}(\hat{\zeta})$	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
$\sigma = 0.5$								
$\widehat{\mathrm{se}}(\hat{\alpha})$	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
$\widehat{\mathrm{se}}(\hat{\beta})$	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
$\widehat{\mathrm{se}}(\hat{\theta})$	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
$\widehat{\mathrm{se}}(\hat{\zeta})$	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
$\sigma = 1$								
$\widehat{\mathrm{se}}(\hat{\alpha})$	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
$\widehat{\mathrm{se}}(\hat{\beta})$	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
$\widehat{\mathrm{se}}(\hat{\theta})$	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
$\widehat{\mathrm{se}}(\hat{\zeta})$	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

†For each parameter the numbers in the first row are the estimated standard errors and the numbers in the second row are the Monte Carlo standard errors.

(c) the weighted least squares mean function estimates, Section 2.5;
(d) the updated variance function estimates, Section 2.6;
(e) 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;
(f) the infeasible variance function estimates when the means are known.

The results are given in Tables 2 and 3, which display the root-mean-squared errors RMSE of various quantities, for example 1 and example 2 respectively. For example, for $\boldsymbol{\alpha}$ RMSE is just $\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|$ (the vectors are normalized to have unit norms) and for m_μ RMSE is $\sqrt{[\Sigma_{t=1}^{50}\{\hat{m}_\mu(x_t) - m_\mu(x_t)\}^2/50]}$, where $(x_1, \ldots, x_{50})$ are equally spaced grid points on an interval with lower bound the 0.01-quantile of the sampled $\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}$ 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 on the estimation of the variance, which is 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 more difficult to estimate than means, and the small sample size effects are

Table 5. Standard errors for the second-stage estimators in example 2 of Section 4.2†

$\sigma=0.2$								
$\widehat{\text{se}}(\hat{\alpha})$	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
$\widehat{\text{se}}(\hat{\beta})$	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
$\widehat{\text{se}}(\hat{\theta})$	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
$\widehat{\text{se}}(\hat{\zeta})$	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
$\sigma=0.5$								
$\widehat{\text{se}}(\hat{\alpha})$	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
$\widehat{\text{se}}(\hat{\beta})$	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
$\widehat{\text{se}}(\hat{\theta})$	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
$\widehat{\text{se}}(\hat{\zeta})$	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
$\sigma=1$								
$\widehat{\text{se}}(\hat{\alpha})$	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
$\widehat{\text{se}}(\hat{\beta})$	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
$\widehat{\text{se}}(\hat{\theta})$	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
$\widehat{\text{se}}(\hat{\zeta})$	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

†For each parameter the numbers in the first row are the estimated standard errors and the numbers in the second row are the Monte Carlo standard errors.

clearly seen, with the updated variance function and parameter estimates dominating the initial estimates, and not as efficient as the infeasible version.

In Figs 2 and 3, we show 20 randomly chosen estimated link functions for the two examples, where the true non-parametric link functions are shown as the full grey 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. For brevity, we consider the second-stage estimators only. 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 that are reported in Tables 4 and 5 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 that are dedicated to the parametric part, and a non-overlapping set of covariates that are dedicated to the non-parametric 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; for example, set $\boldsymbol{\alpha}=(\alpha^{\mathrm{T}},\mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}=(\mathbf{0}^{\mathrm{T}},\beta^{\mathrm{T}})^{\mathrm{T}}$, where α and β are non-zero 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 smoothly clipped absolute deviation can be used to estimate the partition. Essentially, equation (5) is penalized in the usual way to obtain oracle estimates of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Following this, equation (7) would have standard penalties for $(\boldsymbol{\theta}, \boldsymbol{\zeta})$. Under standard conditions, such methods will be oracle, with the penalization parameters estimated by the Akaike information criterion or the Bayesian information criterion. We shall pursue the details of such an approach in the future.

There are many interesting possible extensions of our models (2) and (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 models (2) and (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 to discuss efficiency in this case. However, there are considerable issues with dimensionality, multivariate kernel functions and bandwidth selection. We agree with the sentiment that was 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 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 $R = |Y - m_\mu(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}) - \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}|$ and, working with R, the model can be thought of as

$$E(R|\mathbf{X}) = g\{m_v(\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}) + \mathbf{X}^{\mathrm{T}}\boldsymbol{\zeta}\}, \tag{17}$$

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

where κ is unknown. This is a case where the variance is proportional to a known function of the mean, which is 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.

(a) Except for the non-parametric flavour due to $m_v(\cdot)$, models (17) and (18) can be thought of as generalized linear models. The overwhelming practice in statistics for such models is to use the variance function for weighting only, but not to try to exploit it to obtain further asymptotic efficiency. It is easy to implement this weighting, and to derive an analogue of theorem 2. We have done so, but the promised asymptotic gains in efficiency for estimating $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ were not realized in our simulations.

(b) 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.

(c) Our methods are practical, theoretically justified and clearly work well in simulations and examples.

Acknowledgements

We thank the Associate Editor and two 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 Ministry of Education tier 1 grant. Liang's research was partially supported by National Science Foundation grants DMS-1007167 and DMS-1207444 and by award 11228103, given by the National Natural Science Foundation of China. Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030).

Appendix A

A.1. Assumptions

Assumption 1.

(a) The density function $f_\alpha(u)$ of $\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}$ is bounded away from zero and is continuously differentiable on $(\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}, \mathbf{X}\in\Omega_X)$ 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 neighbourhood of $\boldsymbol{\alpha}_0$.
(b) The function $m_\mu(\cdot)$ is twice continuously differentiable.
(c) The kernel K is a bounded and symmetric probability density function, satisfying

$$\int_{-\infty}^{\infty} u^2\,K(u)\,\mathrm{d}u\neq 0,$$

$$\int_{-\infty}^{\infty} |u|^l\,K(u)\,\mathrm{d}u<\infty, \qquad\qquad l=1,2,3.$$

(d) The matrix Q_ω is positive definite.

Assumption 2.

(a) The density function $f_\theta(t)$ of $\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}$ is bounded away from zero and is continuously differentiable on $(\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}, \mathbf{X}\in\Omega_X)$. Furthermore, $f_\theta(t)$ is uniformly continuous for $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}_0$.
(b) The functions $m_v(\cdot)$ and $g(\cdot)$ are twice continuously differentiable.
(c) $E\{g^{(1)}(\cdot)d\tilde{\Lambda}_\mu|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}=t\}$, $E\{g^{(1)}(\cdot)|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}=t\}$ and $E\{g^{(1)}(\cdot)\Lambda_\mu|\mathbf{X}^{\mathrm{T}}\boldsymbol{\theta}=t\}$ are continuous functions.
(d) The matrix Σ_ϑ defined at expression (11) is positive definite.

A.2. Proof of theorem 1 for the unweighted estimators

The results in theorem 1 for the unweighted estimator follow easily by using the same arguments as in the proof of theorem 4 in Carroll *et al.* (1997).

A.3. Proof of equations (13) and (15) in theorem 2

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

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

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

$$\hat{R}_i-R_i=|Y_i-\{\hat{m}_\mu(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\alpha}})+\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\}|-|Y_i-\{m_\mu(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha})+\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}\}|$$

$$
\begin{aligned}
&= -\hat{S}_i(I_{(\varepsilon_i>0)} - I_{(\varepsilon_i\leqslant 0)}) + \int_0^{\hat{S}_i} (I_{(\varepsilon_i<s)} - I_{(\varepsilon_i\leqslant 0)})\,\mathrm{d}s \\
&= -d_i\tilde{\Lambda}_{i\mu}^{\mathrm{T}}\mathbf{J}_\omega(\hat{\omega}-\omega) + d_i n^{-1}\sum_{j=1}^{n} K_h(U_j-U_i)\frac{\varepsilon_j}{f_\alpha(U_i)} + \int_0^{\hat{S}_i} (I_{(\varepsilon_i<s)} - I_{(\varepsilon_i\leqslant 0)})\,\mathrm{d}s + o_p(n^{-1/2}). \qquad (19)
\end{aligned}
$$

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

A.3.1. Step 1

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

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

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

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(\hat{R}_i - R_i)(1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)} + n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(R_i - g_i)(1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)} \\
&\qquad + n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)[g_i - g\{\hat{a}_0 + \hat{a}_1(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t) + \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\}](1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} \\
&\qquad \times g^{(1)}\{\hat{a_0} + \hat{a}_1(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t) + \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\} + o(n^{-1/2}). \qquad (20)
\end{aligned}
$$

It follows from equation (19) that the first term in equation (20) is

$$
\begin{aligned}
&-n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)} d_i\tilde{\Lambda}_{i\mu}^{\mathrm{T}}\mathbf{J}_\omega(\hat{\omega}-\omega) \\
&\qquad + n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)} d_i n^{-1}\sum_{j=1}^{n} K_h(U_j-U_i)\varepsilon_j/f_\alpha(U_i) \\
&\qquad + n^{-1}\sum_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(1, (\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)} \int_0^{\hat{S}_i} (I_{(\varepsilon_i<s)} - I_{(\varepsilon_i\leqslant 0)})\,\mathrm{d}s + o_p(n^{-1/2}). \qquad (21)
\end{aligned}
$$

For the second term in expression (21), it 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 consider only the first component in the 2×1 vector, since the other is similarly $o_p(n^{-1/2})$. In this case, with $d_i = \mathrm{sgn}(\epsilon_i)$, we want to analyse

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

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

$$
\mathrm{var}(A_n) = n^{-4}\sum_{i,j,p=1}^{n}\sum_{l\neq j} E\{d_i\epsilon_j d_p\epsilon_l\,K_h(U_i-t)\,K_h(U_j-U_i)\,K_h(U_p-t)\,K_h(U_l-U_p)\}.
$$

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

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

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

$$
B_n = c(nh)^{-2}\int K^2(z_1)\,K^2(z_2)\,f_U(t-z_1h)\,f_U(t-z_2h)\,\mathrm{d}z_1\,\mathrm{d}z_2 = O\{(nh)^{-2}\}.
$$

This means that $\mathrm{var}(A_n)=o(n^{-1})$ as long as $nh^2\to\infty$, which follows from our assumptions. Analogously we can prove that the third term in expression (21) is also of order $o_p(n^{-1/2})$.

The second term in expression (20) equals $n^{-1}\Sigma_{i=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}-t)(1,(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}}-t)/h)^{\mathrm{T}} g_i^{(1)}\delta_i$, whereas the third term in expression (20) can be further expressed as

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

A combination of these arguments yields that

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

from which equation (13) follows.

A.3.2. Step 2

Set $\hat{g}_i:=g\{\hat{m}_v(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\zeta}})+\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\}$, $\hat{g}_i^{(1)}:=g^{(1)}\{\hat{m}_v(\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\zeta}})+\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\zeta}}\}$, and let $\hat{\Lambda}_{iv}$ be $\Lambda_{i,v}$ evaluated at $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\zeta}})$. Minimization of expression (7) is equivalent to solving $\Sigma_{i=1}^{n}\hat{\mathbf{J}}_{\vartheta}^{\mathrm{T}}\hat{\Lambda}_{iv}(\hat{R}_i-\hat{g}_i)\hat{g}_i^{(1)}=0$. Now, $\Sigma_{i=1}^{n}\hat{\mathbf{J}}_{\vartheta}^{\mathrm{T}}\hat{\Lambda}_{iv}(\hat{R}_i-\hat{g}_i)\hat{g}_i^{(1)}$ is

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

This is the sum of 10 terms given by

$$
\begin{aligned}
&\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}(\hat{\Lambda}_{iv}-\Lambda_{iv})(\hat{R}_i-R_i)g_i^{(1)}+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}(\hat{\Lambda}_{iv}-\Lambda_{iv})(\hat{R}_i-R_i)(\hat{g}_i^{(1)}-g_i^{(1)})+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}(\hat{\Lambda}_{iv}-\Lambda_{iv})(R_i-g_i)\hat{g}_i^{(1)}\\
&\qquad+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}(\hat{\Lambda}_{iv}-\Lambda_{iv})(\hat{g}_i-g_i)\hat{g}_i^{(1)}+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(\hat{R}_i-R_i)(\hat{g}_i^{(1)}-g_i^{(1)})+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(R_i-g_i)(\hat{g}_i^{(1)}-g_i^{(1)})\\
&\qquad+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(g_i-\hat{g}_i)(\hat{g}_i^{(1)}-g_i^{(1)})+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(\hat{R}_i-R_i)g_i^{(1)}+\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(R_i-g_i)g_i^{(1)}\\
&\qquad-\sum_{i=1}^{n}\mathbf{J}_{\vartheta}^{\mathrm{T}}\Lambda_{iv}(\hat{g}_i-g_i)g_i^{(1)}+o_p(n^{1/2})\\
&\qquad\triangleq\sum_{k=1}^{10}I_k+o_p(n^{1/2}).
\end{aligned}
\tag{22}
$$

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

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

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

$$\begin{aligned}
I_{10} = &- \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} \Lambda_{iv}^{\mathrm{T}} \mathbf{J}_\vartheta g_i^{(1)2} (\hat{\vartheta} - \vartheta) - \sum_{i=1}^{n} \frac{\mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} g_i^{(1)2}}{f_\theta(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}) E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} n^{-1} \sum_{j=1}^{n} K_h(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\theta}) g_j^{(1)} \delta_j \\
&+ \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} g_i^{(1)2} \left(\frac{E\{g^{(1)2}(\cdot)\Lambda_v|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}}{E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} \right)^{\mathrm{T}} \mathbf{J}_\vartheta (\hat{\vartheta} - \vartheta) \\
&+ \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} g_i^{(1)2} \left(\frac{E\{g^{(1)}(\cdot) d\tilde{\Lambda}_\mu|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}}{E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} \right)^{\mathrm{T}} \mathbf{J}_\omega (\hat{\omega} - \omega) + o_p(n^{1/2}) \\
= &- \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} \tilde{\Lambda}_{iv}^{\mathrm{T}} \mathbf{J}_\vartheta g_i^{(1)2} (\hat{\vartheta} - \vartheta) - \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \delta_i g_i^{(1)} \frac{E(\Lambda_{iv} g_i^{(1)2}|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})}{E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} \\
&+ \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} g_i^{(1)2} \left(\frac{E\{g^{(1)}(\cdot) d\tilde{\Lambda}_\mu|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}}{E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} \right)^{\mathrm{T}} \mathbf{J}_\omega (\hat{\omega} - \omega) + o_p(n^{1/2}) \\
= &- \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} \tilde{\Lambda}_{iv}^{\mathrm{T}} \mathbf{J}_\vartheta g_i^{(1)2} (\hat{\vartheta} - \vartheta) - \sum_{i=1}^{n} \mathbf{J}_\vartheta^{\mathrm{T}} \delta_i g_i^{(1)} \frac{E(\Lambda_{iv} g_i^{(1)2}|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta})}{E\{g^{(1)2}(\cdot)|\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\theta}\}} \\
&+ n \mathbf{J}_\vartheta^{\mathrm{T}} E(\Lambda_v g^{(1)} d\tilde{\Lambda}_\mu) \mathbf{J}_\omega (\hat{\omega} - \omega) + o(n^{1/2}).
\end{aligned} \tag{23}$$

We now deal with the term I_8. We have

$$\begin{aligned}
n^{-1} \sum_{i=1}^{n} \mathbf{J}_\omega^{\mathrm{T}} \Lambda_{iv} (\hat{R}_i - R_i) g_i^{(1)} = &-n^{-1} \sum_{i=1}^{n} g_i^{(1)} \mathbf{J}_\omega^{\mathrm{T}} \Lambda_{iv} d_i \tilde{\Lambda}_{i\mu}^{\mathrm{T}} \mathbf{J}_\omega (\hat{\omega} - \omega) + n^{-2} \sum_{i,j} K_h(U_j - U_i) \varepsilon_j g_i^{(1)} \mathbf{J}_\omega^{\mathrm{T}} \Lambda_{iv} d_i / f_\alpha(U_i) \\
&+ n^{-1} \sum_{i=1}^{n} g_i^{(1)} \mathbf{J}_\omega^{\mathrm{T}} \Lambda_{iv} \int_0^{\hat{S}_i} (I_{\{\varepsilon \leqslant s\}} - I_{\{\varepsilon \leqslant 0\}}) \, \mathrm{d}s.
\end{aligned} \tag{24}$$

For the second term in equation (24), 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 $Q_i = g_i^{(1)} \mathbf{J}_\vartheta^{\mathrm{T}} \Lambda_{iv} / f_\alpha(U_i)$. 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

$$\mathbf{S}_n = n^{-1/2} \sum_{i=1}^{n} \varepsilon_i f_\alpha(U_i) \, E(Qd|U_i) + o_p(1) = O_p(1). \tag{25}$$

It suffices to show that

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

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

$$\begin{aligned}
\mathrm{var}(\mathbf{T}_n) &= n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l \neq j} E[\varepsilon_i \varepsilon_k \{K_h(U_j - U_i) Q_j d_j - f_\alpha(U_i) E(Qd|U_i)\} \\
&\quad \times \{K_h(U_l - U_k) Q_l d_l - f_\alpha(U_k) \, E(Qd|U_k)\}] \\
&= n^{-3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l \neq j} E[G_i \{K_h(U_j - U_i) Q_j d_j - f_\alpha(U_i) E(Qd|U_i)\} \\
&\quad \times \{K_h(U_l - U_i) Q_l d_l - f_\alpha(U_i) E(Qd|U_i)\}].
\end{aligned}$$

When $j \neq k$,

$$E[G_i \{K_h(U_j - U_i) Q_j d_j - f_\alpha(U_i) \, E(Qd|U_i)\} \{K_h(U_l - U_i) Q_l d_l - f_\alpha(U_i) \, E(Qd|U_i)\}] = O(h^4) = o(1).$$

Hence,

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

For the third term in equation (24), $\hat{S}_i = O_p\{(nh)^{-1/2}\}$. For any constant $C > 0$ (and similarly for $C < 0$), by a direct calculation of the mean and variance, we have

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

where F is the conditional distribution function of ε and f is the corresponding density. This shows that the third term of equation (24) is also $o_p(n^{-1/2})$.

As a consequence,

$$I_8 = -n\mathbf{J}_\omega^{\mathrm{T}}\, E(g^{(1)} \Lambda_v d \tilde{\Lambda}_\mu^{\mathrm{T}}) \mathbf{J}_\omega(\hat{\omega} - \omega) + \sum_{i=1}^{n} \varepsilon_i f_\alpha(U_i)\, E(Qd|U_i) + o_p(n^{1/2}). \tag{26}$$

A combination of equations (23) and (26) and expression (22) yields that

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

It follows that

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

completing the proof.

A.4. Proof of theorem 1 for the weighted estimators

For the weighted estimators, under the conditions of theorem 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

$$\mathcal{H}(U) = \frac{E(\mathbf{X}/g^2|U)}{E(1/g^2|U)},$$

and $\kappa = m_\mu^{(1)}(U)\{\mathbf{X} - \mathcal{H}(U)\}$. In addition, let

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

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

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

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

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

It is easy to verify that the expectations of these partial derivatives are 0. Accordingly,

$$\begin{aligned} \|\Psi(m_x^*, m_\mu^*, \kappa^*, g^*; \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, \mathbf{X}) - \Psi(\mathcal{H}, m_\mu, \kappa, g; \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, \mathbf{X}) \\ - D(\mathcal{H}^* - \mathcal{H}, m_\mu^* - m_\mu, \kappa^* - \kappa, g^* - g; \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, \mathbf{X})\| \\ = O(\|\mathcal{H}^* - \mathcal{H}\|^2 + \|m_\mu^* - m_\mu\|^2 + \|\kappa^* - \kappa\|^2 + \|g^* - g\|^2), \end{aligned} \tag{27}$$

where $\|\cdot\|$ denotes the Sobolev norm, i.e. the supremum norm of the function itself as well as its derivatives. Equation (27) is Newey's assumption 5.1(i). It is also noteworthy that his assumption 5.2 holds by the expression of $D(\cdot, \cdot, \cdot, \cdot; \boldsymbol{\alpha}, \boldsymbol{\beta}, Y, \mathbf{X})$. Moreover, the result

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

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 $u \in \{\mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}, \mathbf{X} \in \Omega_X\}$:

$$\begin{aligned} \hat{m}_\mu(u) - m_\mu(u) &= o_p(n^{-1/4}), \\ \hat{m}_\mu^{(1)}(u) - m_\mu^{(1)}(u) &= o_p(n^{-1/4}), \\ \hat{\mathcal{H}}(u) - \mathcal{H}(u) &= o_p(n^{-1/4}), \\ \hat{g}(u) - g(u) &= o_p(n^{-1/4}), \\ \hat{g}^{(1)}(u) - g^{(1)}(u) &= o_p(n^{-1/4}). \end{aligned}$$

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

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

$$0 = \sum_{i=1}^n \Psi(\mathcal{H}, m_\mu, \kappa, g; \boldsymbol{\alpha}, \boldsymbol{\beta}, Y_i, \mathbf{X}_i). \tag{28}$$

Furthermore, it can be seen that the solution to equation (28) has the same limit distribution as described in the statement of theorem 1. Hence, we complete the proof for asymptotic normality.

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

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

For any given function q of $\mathbf{X}$, it can be shown that the nuisance tangent space $\mathcal{P}$, for the three nuisance parameters, $q_X(\mathbf{x})$, $p_\epsilon(\epsilon)$ and $m_\mu(U)$, is $\{q(\mathbf{X}) : E(q) = 0, E(\epsilon q) \text{ is a function of } \mathbf{X} \text{ only}\}$. Furthermore, the orthogonal component of $\mathcal{P}$ is $\mathcal{P}^\perp = \{\epsilon q(\mathbf{X})/g^2 : E(q/g^2|U) = 0\}$.

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

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

with

$$\tilde{\mathbf{X}} = \mathbf{X} - \frac{E(\mathbf{X}/g^2|U)}{E(1/g^2|U)}.$$

It can be seen that $S_{\text{eff}} \in \mathcal{P}^{\perp}$.

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

$$\begin{aligned} E\left\{ (S_\omega - S_{\text{eff}})^{\mathrm{T}} \frac{\epsilon q(\mathbf{X})}{g^2} \right\} = \mathbf{J}_\omega E \Biggl[-\begin{pmatrix} m_\mu^{(1)}(U)\mathbf{X} \\ \mathbf{X} \end{pmatrix}^{\mathrm{T}} \frac{q}{g^2} E\left\{ \frac{\epsilon\, p_\epsilon^{(1)}(\epsilon)}{p_\epsilon(\epsilon)} \right\} - \frac{1}{g^2} \Biggl\{ \begin{pmatrix} m_\mu^{(1)}(U)\mathbf{X} \\ \mathbf{X} \end{pmatrix}^{\mathrm{T}} q \\ - E\begin{pmatrix} m_\mu^{(1)}(U) \dfrac{E(\mathbf{X}/g^2|U)}{E(1/g^2|U)} \\ \dfrac{E(\mathbf{X}/g^2|U)}{E(1/g^2|U)} \end{pmatrix}^{\mathrm{T}} q \Biggr\} \Biggr]. \end{aligned}$$

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

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

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

A.5. Proof of equations (14) and (16) in theorem 2

The proof of equations (14) and (16) is identical to that of equations (13) and (15), except that $\hat{\omega}_{\text{WLS}}$ replaces $\hat{\omega}$ everywhere, and we apply the asymptotic expansion of $n^{1/2}(\hat{\omega}_{\text{WLS}} - \omega)$ given in theorem 1. Routine calculations yield theorem 2.

References

Bickel, P. (1978) Using residuals robustly, I: tests for heteroscedasticity, nonlinearity. *Ann. Statist.*, **6**, 266–291.

Bickel, P. J., Klaasen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Box, G. and Hill, W. (1974) Correcting inhomogeneity of variance with power transformation weighting. *Technometrics*, **16**, 385–389.

Box, G. and Meyer, D. (1986) An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Cai, T. T. and Wang, L. (2008) Adaptive variance function estimation in heteroscedastic nonparametric regression. *Ann. Statist.*, **36**, 2025–2054.

Carroll, R. J. (1982) Adapting for heteroscedasticity in linear models. *Ann. Statist*, **10**, 1224–1233.

Carroll, R. J. (2003) Variances are not always nuisance parameters. *Biometrics*, **59**, 211–220.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997) Generalized partially linear single-index models. *J. Am. Statist. Ass.*, **92**, 477–489.

Carroll, R. J. and Härdle, W. (1989) Second order effects in semiparametric weighted least squares regression. *Statistics*, **2**, 179–186.

Carroll, R. J. and Ruppert, D. (1982) Robust estimation in heteroscedasticity linear models. *Ann. Statist.*, **10**, 429–441.

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Davidian, M. and Carroll, R. J. (1987) Variance function estimation. *J. Am. Statist. Ass.*, **82**, 1079–1091.

Davidian, M., Carroll, R. J. and Smith, W. (1988) Variance functions and the minimum detectable concentration in assays. *Biometrika*, **75**, 549–556.

Fuller, W. and Rao, J. (1978) Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Statist.*, **6**, 1149–1158.

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

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

Knight, K. (1998) Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.*, **26**, 755–770.

Liang, H., Liu, X., Li, R. and Tsai, C. L. (2010) Estimation and testing for partially linear single-index models. *Ann. Statist.*, **38**, 3811–3836.

Ma, Y., Chiou, J.-M. and Wang, N. (2006) Efficient semiparametric estimator for heteroscedastic partially linear models. *Biometrika*, **93**, 75–84.

Ma, Y. and Zhu, L. (2013a) Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *J. R. Statist. Soc.* B, **75**, 305–322.

Ma, Y. and Zhu, L. (2013b) Efficient estimation in sufficient dimension reduction. *Ann. Statist.*, **41**, 250–268.

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

Newey, W. K. (1994) The asymptotic variance of semiparametric estimators. *Econometrica*, **62**, 1349–1382.

Teschendorff, A. E. and Widschwendter, M. (2012) Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**, 1487–1494.

Thomas, L., Stefanski, L. A. and Davidian, M. (2012) Measurement error model methods for bias reduction and variance estimation in logistic regression with estimated variance predictors. *Technical Report.* North Carolina State University, Raleigh.

Van Keilegom, I. and Wang, L. (2010) Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electron. J. Statist.*, **4**, 133–160.

Wang, J.-L., Xue, L., Zhu, L. and Chong, Y. S. (2010) Estimation for a partial-linear single-index model. *Ann. Statist.*, **38**, 246–274.

Western, B. and Bloome, D. (2009) Variance function regressions for studying inequality. *Sociol. Methodol.*, **39**, 293–326.

Xia, Y. (2008) A multiple-index model and dimension reduction. *J. Am. Statist. Ass.*, **103**, 1631–1640.

Xia, Y. C. and Härdle, W. (2006) Semi-parametric estimation of partially linear single-index models. *J. Multiv. Anal.*, **97**, 1162–1184.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space (with discussion). *J. R. Statist. Soc.* B, **64**, 363–410.

Yu, Y. and Ruppert, D. (2002) Penalized spline estimation for partially linear single-index models. *J. Am. Statist. Ass.*, **97**, 1042–1054.