ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2014)
76, *Part* 5, *pp.* 885–901

On estimation efficiency of the central mean subspace

Yanyuan Ma

Texas A&M University, College Station, USA

and Liping Zhu

Shanghai University of Finance and Economics, People's Republic of China

[Received November 2012. Revised July 2013]

Summary. We investigate the estimation efficiency of the central mean subspace in the framework of sufficient dimension reduction. We derive the semiparametric efficient score and study its practical applicability. Despite the difficulty caused by the potential high dimension issue in the variance component, we show that locally efficient estimators can be constructed in practice. We conduct simulation studies and a real data analysis to demonstrate the finite sample performance and gain in efficiency of the proposed estimators in comparison with several existing methods.

Keywords: Dimension reduction; Estimating equations; Non-parametric regression; Semiparametric efficiency; Sliced inverse regression

1. Introduction

Data arising from modern sciences often contain a large amount of covariates. The estimation of the central mean subspace is a popular approach of reducing the covariate dimension which allows us to study the relationship between the response and the covariates without relying on a correct mean model (Cook and Li, 2002). Specifically, the approach identifies directions that are represented by the columns of a matrix β, so that the mean of the response Y relates to the covariate vector $\mathbf{x}$ only through $\beta^{\mathrm{T}}\mathbf{x}$. In other words, the conditional mean $E(Y|\mathbf{x})$ is assumed to be a function of $\beta^{\mathrm{T}}\mathbf{x}$ only, although the function itself is left unspecified. When β contains only one column, this reduces to the familiar single-index model. When β contains more than one column, it is sometimes also referred to as a multi-index model (Xia *et al.*, 2002; Xia, 2008).

Most classical dimension reduction methods are based on inverse regression, pioneered by sliced inverse regression (Li, 1991). For identifying the central mean subspace, Li and Duan (1989) suggested ordinary least squares methods, and Li (1992) and Cook and Li (2002) proposed the principal Hessian direction (PHD) method. Following the idea of Cook and Ni (2005, 2006), Yoo and Cook (2007) proposed weighted alternating least squares to improve the usual ordinary least squares estimator when Y is multivariate. These methods assume either the linearity condition alone, where $E(\mathbf{x}|\beta^{\mathrm{T}}\mathbf{x})$ is a linear function of $\beta^{\mathrm{T}}\mathbf{x}$, or the constant variance condition in addition, where $\mathrm{cov}(\mathbf{x}|\beta^{\mathrm{T}}\mathbf{x})$ is a constant matrix. A second main idea in identifying the central mean subspace is to estimate the unknown mean function and the central mean

Address for correspondence: Liping Zhu, School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai, 200433, People's Republic of China.
E-mail: zhu.liping@mail.shufe.cn

 1369–7412/14/76885

subspace simultaneously. Methods in this class naturally involve non-parametric regression. Continuous covariate distributions are typically assumed and iteration procedures are often used in these methods. The representative work is the minimum average variance estimator (MAVE) method (Xia *et al.*, 2002). Recently, Ma and Zhu (2012) proposed a third semiparametric approach, where they used semiparametric techniques to estimate the central mean subspace through solving estimating equations. This approach does not rely on any moment or distributional conditions on the covariates and hence is applicable in a wider range. However, Ma and Zhu (2012) did not further study inference or efficiency issues, as are studied here.

The main focus of this paper is to investigate the inference and estimation efficiency issue in recovering the central mean subspace. The general approach that we take is semiparametric and geometric, as illustrated in Bickel *et al.* (1993) and Tsiatis (2006). We benefit from a simple parameterization that converts the space identification problem into a problem of estimating a finite dimensional parameter in a semiparametric model. This allows us to derive the semiparametric efficient score, which has the potential of reaching the minimum estimation variance bound among all possible consistent estimators. All the derivations are performed without using the linearity condition, or the constant variance condition or the continuity of the covariates. We study the efficient score expression and point out the implementational difficulties as well as provide strategies to circumvent this difficulty in practice. Efficiency study has also been carried out in some simpler models (Ma *et al.*, 2006) and under stronger conditions (Delecroix *et al.*, 2003). As far as we know, this is the first work on estimation inference and efficiency of the central mean subspace model in its general form without imposing any additional model assumptions.

The rest of this paper is as follows. In Section 2, we derive the efficient score and the theoretical form of the efficient estimator. We study the practical difficulty of the implementation of the efficient estimator in Section 3 and propose several locally efficient estimators and show their asymptotic properties. Simulation studies are conducted in Section 4 to demonstrate the practical gain in efficiency and the method is implemented in an example in Section 5. We finish the paper with a brief discussion in Section 6. All the technical derivations are given in Appendix A and in an on-line supplement.

2. Semiparametric formulation and efficient score

Let $\mathbf{x}$ be a $p \times 1$ covariate vector and Y be a univariate response variable. Cook and Li (2002) introduced the concept of the central mean subspace where the conditional mean of the response is assumed to depend on a few linear combinations of the covariates. Specifically, assume that $\boldsymbol{\beta}$ satisfies

$$E(Y|\mathbf{x}) = E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}); \tag{1}$$

then the column space of $\boldsymbol{\beta}$ is defined as a mean dimension reduction subspace. The central mean subspace, which is denoted by $\mathcal{S}_{E(Y|\mathbf{x})}$, is subsequently defined as the intersection of all mean dimension reduction subspaces if the intersection itself is a mean dimension reduction subspace. The conditional mean model assumes that $\mathbf{x}$ contributes to the conditional mean of Y only through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. The main interest for this model is typically in estimating $\mathcal{S}_{E(Y|\mathbf{x})}$ or, equivalently, a basis matrix $\boldsymbol{\beta}$ which spans $\mathcal{S}_{E(Y|\mathbf{x})}$. The central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$ has a well-known invariance property (Cook and Li, 2002), which allows us to assume that, without loss of generality, the covariate vector $\mathbf{x}$ satisfies $E(\mathbf{x}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{x}) = \mathbf{I}_p$. The dimension of $\mathcal{S}_{E(Y|\mathbf{x})}$, which is denoted by d, is referred to as the structural dimension. Typically, $\mathcal{S}_{E(Y|\mathbf{x})}$ is identified through estimating a full rank matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d}$ that satisfies condition (1) and forming

its d-dimensional column space. Although the central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$ is unique, $\boldsymbol{\beta}$ is not. In fact, for any $d \times d$ full rank matrix $\mathbf{A}$, $\boldsymbol{\beta}\mathbf{A}$ generates the same $\mathcal{S}_{E(Y|\mathbf{x})}$ as $\boldsymbol{\beta}$.

To avoid the problem that many $\boldsymbol{\beta}$-matrices can represent the same central mean subspace $\mathcal{S}_{E(Y|\mathbf{x})}$, we use the parameterization which requires $\boldsymbol{\beta}$ to have its upper $d \times d$ submatrix equal to the identity matrix $\mathbf{I}_d$, whereas the lower $(p-d) \times d$ submatrix is not subject to any constraints. Ma and Zhu (2013) showed that this parameterization successfully converts the problem of identifying $\mathcal{S}_{E(Y|\mathbf{x})}$ to the problem of estimating $\boldsymbol{\beta}$, which contains $p_t = (p-d)d$ free parameters. Estimating $\boldsymbol{\beta}$ in equation (1) then becomes a typical semiparametric estimation problem.

The conditional mean model (1) can be equivalently written as

$$Y = m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) + \varepsilon, \tag{2}$$

where m is an unspecified smooth function and $E(\varepsilon|\mathbf{x}) = 0$. In this formulation, it is easy to see that the likelihood of one random observation $(\mathbf{x}, Y)$ is

$$\eta_1(\mathbf{x})\,\eta_2\{Y - m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \mathbf{x}\},$$

where η_1 captures the marginal density of $\mathbf{x}$, $m(\cdot)$ is the mean function of Y conditional on $\mathbf{x}$ and η_2 is the probability density function of the residual $\varepsilon = Y - E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ conditional on $\mathbf{x}$. Here, η_2 also satisfies $E(\varepsilon|\mathbf{x}) = 0$. Treating η_1, η_2 and m as nuisance parameters and $\boldsymbol{\beta}$ as the parameter of interest, we view the estimation of $\boldsymbol{\beta}$ as a semiparametric problem and derive its corresponding efficient score function. For this, let $\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})$ be the derivative of η_2 with respect to ε; the nuisance tangent space of the model (Ma and Zhu, 2012) is $\Lambda = \Lambda_1 \oplus \Lambda_2 + \Lambda_m$, where

$$\begin{aligned}
\Lambda_1 &= \{f(\mathbf{x}) : \forall f \text{ subject to } F(f) = 0\},\\
\Lambda_2 &= \{f(\varepsilon, \mathbf{x}) : \forall f \text{ subject to } E(f|\mathbf{x}) = 0, E(\varepsilon f|\mathbf{x}) = 0\},\\
\Lambda_m &= \left\{ \frac{\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})}{\eta_2(\varepsilon, \mathbf{x})} h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) : \forall h \right\}.
\end{aligned}$$

Here, '+' represents the sum of two spaces, i.e. the space formed by the linear combination of any elements in either of the two spaces, whereas '$\oplus$' represents the sum of two spaces that are orthogonal to each other. To gain a more helpful form of Λ, we calculate the residual after projecting a function in Λ_m to Λ_2 to obtain

$$\Lambda'_m = \Pi(\Lambda_m|\Lambda_2^{\perp}) = \left\{ \frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) : \forall h \right\},$$

where $\Pi(\Lambda_m|\Lambda_2^{\perp})$ represents the orthogonal projection of Λ_m to $\Lambda_2^{\perp}$. Thus, $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda'_m$. It is easy to verify that the orthogonal complement of Λ is

$$\begin{aligned}
\Lambda^{\perp} &= ([\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}]\varepsilon : \forall \boldsymbol{\alpha}(\mathbf{x}))\\
&= (\{Y - E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}[\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}] : \forall \boldsymbol{\alpha}(\mathbf{x})).
\end{aligned}$$

Let $\mathrm{vecl}(\mathbf{M})$ be the concatenation of the lower $(p-d) \times d$ block of a $p \times d$ matrix $\mathbf{M}$. Straightforward calculation yields the score function

$$\mathbf{S}_{\beta} = -\mathrm{vecl}\left\{ \frac{\mathbf{x}\,\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})\,\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\eta_2(\varepsilon, \mathbf{x})\,\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})} \right\},$$

which we further decompose into

$$\mathbf{S}_\beta = -\mathrm{vecl}\left[\left\{\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} + \frac{\eta'_{2\varepsilon}(\varepsilon,\mathbf{x})}{\eta_2(\varepsilon,\mathbf{x})}\right\}\frac{\mathbf{x}\,\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})}\right]$$
$$+\mathrm{vecl}\left[\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})}\frac{E\{\mathbf{x}E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}{E\{E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}\frac{\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})}\right]$$
$$+\mathrm{vecl}\left(\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})}\left[\mathbf{x}-\frac{E\{\mathbf{x}E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}{E\{E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}\right]\frac{\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})}\right).$$

We can now easily verify that the first component is in Λ_2, the second component is in Λ'_m and the third component is in $\Lambda^{\perp}$. Hence we obtain

$$\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w(\mathbf{x})\} = \mathrm{vecl}\left(\varepsilon\, w(\mathbf{x})\left[\mathbf{x}-\frac{E\{\mathbf{x}\,w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}{E\{w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}\right]\mathbf{m}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right), \tag{3}$$

where $w^{-1}(\mathbf{x}) = E(\varepsilon^2|\mathbf{x})$, $\mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})/\partial(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$. Hypothetically, the efficient estimator can then be obtained from the sample version of $E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w(\mathbf{x})\}] = \mathbf{0}$.

3. Locally efficient and efficient estimation of $\mathcal{S}_{E(Y|\mathbf{x})}$

Implementing an estimation procedure based on $\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w(\mathbf{x})\}$ in equation (3) requires estimation of $m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, its first derivative $\mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, the inverse of the error variance function $w(\mathbf{x})$ and the expectations $E\{w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$ and $E\{\mathbf{x}\,w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$. At a fixed $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ has dimension d, which, by the very purpose of dimension reduction, is a dimension that can be handled in practice. This suggests that $m(\cdot)$, $\mathbf{m}_1(\cdot)$ and $E(\cdot|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ can be estimated via proper non-parametric treatment. For example, $\hat{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\hat{\mathbf{m}}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ at $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0$ can be obtained through a local linear approximation procedure by minimizing

$$\sum_{i=1}^{n}\{Y_i - m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) - \mathbf{m}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\}^2 K_{h_1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0), \tag{4}$$

where K is the multiplication of d univariate kernel functions, denoted by K as well for simplicity, h_1 is a bandwidth and

$$K_{h_1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) = \frac{K\{(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)/h_1\}}{{h_1}^d}.$$

Similarly, to resolve the issue of estimating $E\{w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$ and $E\{\mathbf{x}\,w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$, we can use, for example, the kernel estimators

$$\hat{E}\{w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0\} = \sum_{i=1}^{n} w(\mathbf{x}_i)K_{h_2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\Big/\sum_{i=1}^{n} K_{h_2}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0),$$

and

$$\hat{E}\{\mathbf{x}\,w(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0\} = \sum_{i=1}^{n} \mathbf{x}_i w(\mathbf{x}_i)K_{h_3}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\Big/\sum_{i=1}^{n} K_{h_3}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0). \tag{5}$$

We point out that, if $w(x)$ happens to be a constant, then, under the linearity condition that $E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0) = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0$, no non-parametric estimation is needed. However, even with constant weight, performing non-parametric kernel estimation allows us to remove the linearity condition. The situation is quite different in terms of estimating $w(\mathbf{x})$. Viewing ε^2 as a response variable and $\mathbf{x}$ as a covariate vector, this is exactly the same mean regression

problem where the high dimension of the covariates initiated the original central mean subspace modelling. Thus, its estimation suffers the curse of dimensionality and is to be avoided preferably. To circumvent this difficulty, a helpful compromise is to consider plausible forms for $w(\mathbf{x})$, instead of estimating it as a completely unspecified positive smooth function of $\mathbf{x}$, provided that consistency is retained. In other words, we seek local efficiency in handling the unknown $w(\mathbf{x})$.

3.1. Fixed form of $w(\mathbf{x})$

The simplest treatment is to replace the unknown $w(\mathbf{x})$ by a known positive function $w_1(\mathbf{x})$ that is the best current guess for $w(\mathbf{x})$. For example, $w_1(\mathbf{x})$ can be a constant, reflecting a homoscedastic error conjecture. This yields the estimating equation

$$\sum_{i=1}^{n} \mathrm{vecl}\left(\{Y_i - \hat{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\, w_1(\mathbf{x}_i) \left[\mathbf{x}_i - \frac{\hat{E}\{\mathbf{x}_i\, w_1(\mathbf{x}_i)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}}{\hat{E}\{w_1(\mathbf{x}_i)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}} \right] \hat{\mathbf{m}}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \right) = \mathbf{0}. \tag{6}$$

Of course, without any additional information, it is very unlikely that a guessed $w_1(\mathbf{x})$ can equal or even resemble the true weight function $w(\mathbf{x})$. The fortunate fact is that, whether or not $w_1(\mathbf{x}) = w(\mathbf{x})$, solving equation (6) will always result in a consistent estimator of $\boldsymbol{\beta}$. If it happens that $w_1(\mathbf{x}) = w(\mathbf{x})$, the estimator is also efficient. We state this result in theorem 1.

Theorem 1. Assume that $\hat{\boldsymbol{\beta}}_1$ solves equation (6). Under conditions 1–5 in Appendix A when $n \to \infty$,

$$\sqrt{n}\,\mathrm{vecl}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \to N\{\mathbf{0}, \mathbf{A}_1^{-1}\mathbf{B}_1(\mathbf{A}_1^{-1})^{\mathrm{T}}\}$$

in distribution, where

$$\mathbf{A}_1 = E\left[\frac{\partial \mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_1(\mathbf{x})\}}{\partial \mathrm{vecl}(\boldsymbol{\beta})^{\mathrm{T}}} \right],$$

$$\mathbf{B}_1 = E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_1(\mathbf{x})\}\, \mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_1(\mathbf{x})\}].$$

When $w_1(\mathbf{x}) = w(\mathbf{x})$, $\mathbf{A}_1 = \mathbf{B}_1 = E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\}\, \mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\}]$, and $\hat{\boldsymbol{\beta}}_1$ is efficient.

3.2. Parametric modelling for $w(\mathbf{x})$

A more cautious approach of achieving good efficiency is to propose a model for $w(\mathbf{x})$. In the parametric modelling strategy, we use $w_2(\mathbf{x}, \boldsymbol{\zeta})$ to approximate $w(\mathbf{x})$. Here w_2 has a prespecified form and $\boldsymbol{\zeta}$ is an unspecified parameter vector. In practice, we could estimate $\boldsymbol{\zeta}$ through regressing the squared residuals $\tilde{\varepsilon}_i^2 = \{Y_i - \hat{m}(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{x}_i)\}^2$ onto $\mathbf{x}_i$, for $i = 1, \ldots, n$, where $\hat{m}(\cdot)$ and $\hat{\boldsymbol{\beta}}_1$ are initial estimates obtained from Section 3.1. Write the resulting estimate of $\boldsymbol{\zeta}$ as $\hat{\boldsymbol{\zeta}}$. If the family $\{w_2(\mathbf{x}, \boldsymbol{\zeta}) : \forall \boldsymbol{\zeta}\}$ is sufficiently flexible that it contains the true weights $w(\mathbf{x})$, i.e. there exists $\boldsymbol{\zeta}_0$ such that $w_2(\mathbf{x}, \boldsymbol{\zeta}_0) = w(\mathbf{x})$, then the resulting estimator for $\boldsymbol{\zeta}$ based on solving

$$\sum_{i=1}^{n} \mathrm{vecl}\left(\{Y_i - \hat{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\, w_2(\mathbf{x}_i, \hat{\boldsymbol{\zeta}}) \left[\mathbf{x}_i - \frac{\hat{E}\{\mathbf{x}_i\, w_2(\mathbf{x}_i, \hat{\boldsymbol{\zeta}})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}}{\hat{E}\{w_2(\mathbf{x}_i, \hat{\boldsymbol{\zeta}})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}} \right] \hat{\mathbf{m}}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i) \right) = \mathbf{0} \tag{7}$$

will be efficient. Even if the family $\{w_2(\mathbf{x}, \boldsymbol{\zeta}) : \forall \boldsymbol{\zeta}\}$ does not contain $w(\mathbf{x})$, the resulting estimator for $\boldsymbol{\beta}$ remains consistent. We state this result in theorem 2.

Theorem 2. Assume that $\hat{\boldsymbol{\beta}}_2$ solves equation (7) and $\hat{\boldsymbol{\zeta}}$ is a root-n-consistent estimate of $\boldsymbol{\zeta}$. Under conditions 1–5 in Appendix A, when $n \to \infty$,

$$\sqrt{n}\,\mathrm{vecl}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \to N\{\mathbf{0}, \mathbf{A}_2^{-1}\mathbf{B}_2(\mathbf{A}_2^{-1})^{\mathrm{T}}\}$$

in distribution, where

$$\mathbf{A}_2 = E\left[\frac{\partial \mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_2(\mathbf{x}, \boldsymbol{\zeta})\}}{\partial \mathrm{vecl}(\boldsymbol{\beta})^{\mathrm{T}}}\right],$$

$$\mathbf{B}_2 = E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_2(\mathbf{x}, \boldsymbol{\zeta})\}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w_2(\mathbf{x}, \boldsymbol{\zeta})\}].$$

When $w_2(\mathbf{x}, \boldsymbol{\zeta}) = w(\mathbf{x})$, $\mathbf{A}_2 = \mathbf{B}_2 = E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x}, Y, \boldsymbol{\beta}, w(\mathbf{x})\}]$, and $\hat{\boldsymbol{\beta}}_2$ is efficient.

Remark 1. We would like to point out a curious observation that, in estimating $\boldsymbol{\beta}$ from equation (7), the variability of estimating $\boldsymbol{\zeta}$ does not have an effect on the variability of $\hat{\boldsymbol{\beta}}_2$. As long as $\hat{\boldsymbol{\zeta}}$ is root n consistent, using $\hat{\boldsymbol{\zeta}}$ is equally effective as using $\boldsymbol{\zeta}$ as far as the asymptotic variance of $\hat{\boldsymbol{\beta}}_2$ is concerned. This seems to suggest that one should use a rich model $w_2(\mathbf{x}, \boldsymbol{\zeta})$ so that the chance of capturing the truth $w(\mathbf{x})$ is large. Although this is so theoretically, in practice, with a finite sample size, a large model indicates a large length of the parameter vector $\boldsymbol{\zeta}$, and its estimation could be unstable owing to numerical issues.

Remark 2. When the family $\{w_2(\mathbf{x}, \boldsymbol{\zeta}) : \forall \boldsymbol{\zeta}\}$ contains the truth $w(\mathbf{x})$, i.e. there exists $\boldsymbol{\zeta}_0$ so that $w_2(\mathbf{x}, \boldsymbol{\zeta}_0) = w(\mathbf{x})$, performing a regression of the residual squares on the covariates for estimation will yield a root-n-consistent estimator of $\boldsymbol{\zeta}_0$. In contrast, when the family does not contain the truth, the limit of $\hat{\boldsymbol{\zeta}}$ by using the same procedure will be a $\boldsymbol{\zeta}$-value which minimizes the distance $E[\{w^{-1}(\mathbf{x}) - {w_2}^{-1}(\mathbf{x}, \boldsymbol{\zeta})\}^2]$ among all possible $\boldsymbol{\zeta}$-values.

Remark 3. In practice, it is not known whether $\{w_2(\mathbf{x}, \boldsymbol{\zeta}) : \forall \boldsymbol{\zeta}\}$ contains $w(\mathbf{x})$. Thus, an alternative approach is through profiling. We obtain $\hat{\boldsymbol{\beta}}_2(\boldsymbol{\zeta})$ via solving equation (7), but with $\boldsymbol{\zeta}$ left unspecified, and then we obtain $\hat{\boldsymbol{\zeta}}$ and subsequently $\hat{\boldsymbol{\beta}}_2(\hat{\boldsymbol{\zeta}})$ through minimizing some measure of the estimated asymptotic variance, say $\mathrm{tr}\{\hat{\mathbf{A}}_2^{-1}\hat{\mathbf{B}}_2(\hat{\mathbf{A}}_2^{-1})^{\mathrm{T}}\}$ as a function of $\boldsymbol{\zeta}$. Because this strategy optimizes the estimated asymptotic variance, hence it is applicable only when the sample size is sufficiently large that the asymptotic variance is a good approximation of the finite sample performance and $\hat{\mathbf{A}}_2^{-1}\hat{\mathbf{B}}_2(\hat{\mathbf{A}}_2^{-1})^{\mathrm{T}}$ approximates well the asymptotic variance $\mathbf{A}_2^{-1}\mathbf{B}_2(\mathbf{A}_2^{-1})^{\mathrm{T}}$ as well.

3.3. Semiparametric modelling for $w(\mathbf{x})$

We can also use a semiparametric modelling strategy to approximate $w(\mathbf{x})$ with an unspecified function $w_3\{\boldsymbol{\xi}(\mathbf{x})\}$. Here the functional form of w_3 is unknown, and $\boldsymbol{\xi}(\mathbf{x})$ has a prespecified form and is of sufficiently low dimension that it is practical to perform a non-parametric smoothing procedure to estimate w_3. For example, a natural choice is $\boldsymbol{\xi}(\mathbf{x}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ if we believe that the same dimension reduction feature for the mean applies to the variance also. This practice was adopted in Härdle *et al.* (1993), although they further assumed that the variance as a function of $\boldsymbol{\xi}$ is known up to a scale, which hence greatly simplifies the problem. A more flexible choice is to assume that $\boldsymbol{\xi}(\mathbf{x}) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$, where $\boldsymbol{\gamma}$ does not have to equal $\boldsymbol{\beta}$. Here, $\boldsymbol{\gamma}$ preferably contains no more than d columns; thus the estimation of $\boldsymbol{\gamma}$ and w_3, based on $\tilde{\varepsilon}_i^2$ and $\mathbf{x}_i$, $i = 1, \ldots, n$, is not more difficult than the original problem of estimating the central mean subspace. We can of course also adopt other forms for $\boldsymbol{\xi}$, such as $\boldsymbol{\xi} = \mathbf{x}^{\mathrm{T}}\mathbf{x}$ or $\boldsymbol{\xi} = (X_1, X_2^2)^{\mathrm{T}}$. When $\boldsymbol{\xi}$ does not contain unknown parameters, the estimation of w_3 can be easily obtained via

$$\hat{w}_3(\boldsymbol{\xi}) = \sum_{i=1}^{n} K_{h_4}(\boldsymbol{\xi}_i - \boldsymbol{\xi}) \Big/ \sum_{i=1}^{n} \tilde{\varepsilon}_i^2 K_{h_4}(\boldsymbol{\xi}_i - \boldsymbol{\xi}). \tag{8}$$

Here and in equation (9), K is the multiplication of $\dim(\boldsymbol{\xi})$ univariate kernel functions, where $\dim(\boldsymbol{\xi})$ is the length of $\boldsymbol{\xi}$ and h_4 is a bandwidth. This also applies if $\boldsymbol{\xi}=\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, where we simply replace $\boldsymbol{\xi}_i$ in the above display with $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i$. When $\boldsymbol{\xi}=\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$ is adopted, we suggest that $\hat{\boldsymbol{\gamma}}$ is obtained from solving

$$\sum_{i=1}^{n}\mathrm{vecl}\left[\left\{\hat{\varepsilon}_i^2-\frac{\sum_{j=1}^{n}\hat{\varepsilon}_j^2K_b(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_j-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_i)}{\sum_{j=1}^{n}K_b(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_j-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_i)}\right\}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}\right]=\mathbf{0}, \tag{9}$$

and then $\hat{w}_3(\boldsymbol{\xi})$ is obtained through equation (8) with $\boldsymbol{\xi}_i$ replaced by $\hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{x}_i$. Here b is a bandwidth. The estimation of $\boldsymbol{\beta}$ is then obtained from solving

$$\sum_{i=1}^{n}\mathrm{vecl}\left(\{Y_i-\hat{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}\hat{w}_3(\boldsymbol{\xi}_i)\left[\mathbf{x}_i-\frac{\hat{E}\{\mathbf{x}_i\hat{w}_3(\boldsymbol{\xi}_i)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}}{\hat{E}\{\hat{w}_3(\boldsymbol{\xi}_i)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i\}}\right]\hat{\mathbf{m}}_1^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\right)=\mathbf{0}. \tag{10}$$

Similarly, when $w_3(\boldsymbol{\xi})$ is a correct model, then the resulting estimator for $\boldsymbol{\beta}$ is efficient. Otherwise, it is still consistent. We state this result in theorem 3.

Theorem 3. Assume that $\hat{\boldsymbol{\beta}}_3$ solves equation (10). Under conditions 1–5 in Appendix A, when $n\to\infty$,

$$\sqrt{n}\,\mathrm{vecl}(\hat{\boldsymbol{\beta}}_3-\boldsymbol{\beta})\to N\{\mathbf{0},\mathbf{A}_3^{-1}\mathbf{B}_3(\mathbf{A}_3^{-1})^{\mathrm{T}}\}$$

in distribution, where

$$\mathbf{A}_3=E\left[\frac{\partial\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w_3(\boldsymbol{\xi})\}}{\partial\mathrm{vecl}(\boldsymbol{\beta})^{\mathrm{T}}}\right],$$

$$\mathbf{B}_3=E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w_3(\boldsymbol{\xi})\}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x},Y,\boldsymbol{\beta},w_3(\boldsymbol{\xi})\}].$$

When $w_3(\boldsymbol{\xi})=w(\mathbf{x})$, $\mathbf{A}_3=\mathbf{B}_3=E[\mathbf{S}_{\mathrm{eff}}\{\mathbf{x},Y,\boldsymbol{\beta},w(\mathbf{x})\}\mathbf{S}_{\mathrm{eff}}^{\mathrm{T}}\{\mathbf{x},Y,\boldsymbol{\beta},w(\mathbf{x})\}]$ and $\hat{\boldsymbol{\beta}}_3$ is efficient.

Remark 4. When the family $\{w_3(\boldsymbol{\xi}):\forall w_3\}$ does not contain the truth $w(\mathbf{x})$, i.e. $w_3(\boldsymbol{\xi})\neq w(\mathbf{x})$ for all w_3, $\hat{w}_3(\boldsymbol{\xi})$ can be viewed as an estimate of the function that minimizes the distance $E[\{w_3^{-1}(\boldsymbol{\xi})-w^{-1}(\mathbf{x})\}^2]$ among all smooth functions of $\boldsymbol{\xi}$.

Remark 5. The decision on whether we should use a fixed form, a parametric model or a semiparametric model of $w(\mathbf{x})$ relies on what is the priority of the practice. If we are willing to sacrifice the performance of the estimator to gain a simple and quick answer, a fixed form of $w(\mathbf{x})$ is a proper choice. However, if estimation efficiency is important and data are sufficiently large, then a semiparametric model is the most reliable choice. A parametric model is in between the two decisions and is suitable when a balance between performance and computation is sought. Generally speaking, modelling $w(\mathbf{x})$ is not very different from any standard variance modelling procedure, except that here we have the added security that a wrong model will not harm the consistency of the estimation of $\boldsymbol{\beta}$, and large estimation variability in the variance model does not cause a loss of efficiency in estimating $\boldsymbol{\beta}$.

3.4. Implementation

Although we have the ability to estimate m, $\mathbf{m}_1$ and $E(\cdot|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, we can also opt not to perform all of the corresponding non-parametric estimations if merely consistency is needed. Similarly to the treatment for $w(\mathbf{x})$, we can also simply propose some models for these quantities and plug

them into equation (3). However, it is important to estimate at least one of m and $E(\cdot|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ to guarantee the consistency. In other words, to ensure consistency, we are free to replace $\mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ by an arbitrary function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, and to choose either to estimate m faithfully which is required in forming ε or to estimate $E\{w(\cdot)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$ and $E\{\mathbf{x}\,w(\cdot)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}$ faithfully. Among all the previous modelling procedures, the choice of selecting $w_3(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$ as the model for $w(\mathbf{x})$ is the most complex in terms of implementation. Hence we outline the implementation details here.

Step 1: we obtain an initial root-n-consistent estimator of $\boldsymbol{\beta}$ through, say, equation (6) or the semiparametric PHD method (Ma and Zhu, 2012). Denote this initial estimator $\tilde{\boldsymbol{\beta}}$.
Step 2: obtain $\hat{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ and $\hat{\mathbf{m}}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)$ from minimizing equation (4) with $\boldsymbol{\beta}$ replaced by $\tilde{\boldsymbol{\beta}}$.
Step 3: form $\hat{\varepsilon}_i = Y_i - \hat{m}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)$. Obtain $\hat{\boldsymbol{\gamma}}$ from solving equation (9). Form $\boldsymbol{\xi}_i = \hat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{x}_i$ and obtain $\hat{w}_3$ from equation (8) at $\boldsymbol{\xi} = \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$.
Step 4: obtain $\hat{E}\{\hat{w}_3(\boldsymbol{\xi}_i)|\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i\}$ and $\hat{E}\{\mathbf{x}_i\hat{w}_3(\boldsymbol{\xi}_i)|\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i\}$ by using equation (5).
Step 5: form equation (10) and solve it by using the Newton–Raphson algorithm to obtain $\hat{\boldsymbol{\beta}}$.

To achieve the asymptotic property that is described in theorem 3, the steps described above do not need to be iterated. However, in practice, iteration is almost always needed to improve the finite sample performance. Of course none of the modelling procedures that were mentioned in Section 3 can guarantee efficiency. The only way to guarantee the efficiency in estimating $\boldsymbol{\beta}$ is to estimate $w(\mathbf{x})$ in a model-free fashion, which is nearly impossible in practice owing to the curse of dimensionality. Although a higher order multivariate kernel can theoretically resolve this issue, it is not feasible unless the sample size is very large.

Our final note on implementation concerns selecting the various bandwidths that are associated with the respective non-parametric regressions. In practice, we have found that the estimation procedure is not very sensitive to the bandwidth. This is not a surprise given that the insensitivity to bandwidth is a universal phenomenon in semiparametric problems. See Van Keilegom and Carroll (2007) for a refined study and Maity *et al.* (2007) for extensive numerical experiments on the bandwidth issue in semiparametric models. The insensitivity is also reflected in the regularity condition 5 in Appendix A, in that a wide range of bandwidths is allowed and all will yield the same first-order asymptotic properties of $\boldsymbol{\beta}$. Because of this insensitivity, we suggest two ways in practical implementation. The easier is to use the sample size and condition 5 to obtain the right order of the bandwidths, and to use the data or initial estimator information to scale them properly. For example, in the data analysis, we simply used the bandwidths at the suitable order by setting $h_1 = cn^{-1/(2+d)}$ and $h_2 = h_3 = h_4 = cn^{-1/(4+d)}$, where c is the average standard deviation of $\mathbf{x}$. A slightly more careful procedure is to use a standard procedure such as cross-validation or to plug in to select the suitable bandwidths. Note that no further scaling is necessary since the classical optimal non-parametric bandwidths automatically satisfy condition 5.

4. Simulation study

We perform two simulation studies to demonstrate the performance of the various locally efficient estimators and compare them with several existing methods including MAVE (Xia *et al.*, 2002) and the PHD method (Li, 1992).

In the first simulation, we set $p = 6$ and $d = 1$, and let $\boldsymbol{\beta} = (1, 0, 1, 1, 1, 1)^{\mathrm{T}}/\sqrt{5}$. The true mean function is $m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + 1)$. The error is normal with the true variance function $\sigma^2(\mathbf{x}) = \{(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^2 + 1\}/2$.

We consider two scenarios for generating $\mathbf{x}$ in this example.

(a) In scenario 1, we generate $\mathbf{x}$ from the multivariate normal distribution, where the covariance between the ith and jth component is $0.5^{|i-j|}$.
(b) In scenario 2, we generate five discrete random variables $X_1^*, \ldots X_5^*$, each with value 1 or -1 with probability 0.5, and one standard normal random variable X_6^*. We then perform a linear transformation on $\mathbf{x}^* = (X_1^*, \ldots, X_6^*)^{\mathrm{T}}$ to obtain $\mathbf{x}$, so that the variance–covariance structure of $\mathbf{x}$ is exactly the same as before. This scenario is designed to demonstrate how well our methods can handle discrete covariates. Note that, if all the components in $\mathbf{x}$ are binary, the problem is not identifiable (Horowitz and Härdle, 1996).

In the second simulation, we consider $d = 2$. We set the two columns of $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 1, 1)^{\mathrm{T}}/\sqrt{6}$ and $\boldsymbol{\beta}_2 = (1, -1, 1, -1, 1, -1)^{\mathrm{T}}/\sqrt{6}$. We form the covariates $\mathbf{x}$ by setting $X_1 = U_1$, $X_2 = U_2 - U_1$, $X_3 = U_3 - U_2$, $X_4 = 2U_4 + U_3 - U_2$, $X_5 = U_5$ and $X_6 = U_5/2 + U_6$, where $U_1, \ldots, U_6$ are independent uniform random variables between $-\sqrt{3}$ and $\sqrt{3}$. This allows correlation between the covariates. The true mean function is $m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1) + (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)^2$, and the error is generated from a normal distribution with mean 0 and true variance function $\sigma^2(\mathbf{x}) = \log\{(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1)^2 + (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_2)^2 + 2\}$.

In both simulations, we run 1000 simulations with sample size $n = 500$. We implemented six different estimators.

(a) The oracle estimator: we obtain the estimator from the score function (3) with a known weight function $w(\mathbf{x}) = \sigma^2(\mathbf{x})^{-1}$. This is the optimal estimator and it serves as a benchmark.

Table 1. Average of estimators, ave, sample estimation standard deviation, sd, average of the estimated standard deviation, $\widehat{\text{sd}}$, and 95% coverage, cvg, of various estimators in simulation 1, scenarios 1 and 2

Method	*Statistic*	*Results for scenario 1*					*Results for scenario 2*				
		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
Truth		0.553	1.106	1.032	0.885	0.590	0.553	1.106	1.032	0.885	0.590
Oracle	ave	0.563	1.129	1.054	0.902	0.601	0.564	1.128	1.052	0.902	0.601
	sd	0.041	0.052	0.048	0.046	0.040	0.043	0.051	0.049	0.047	0.040
	$\widehat{\text{sd}}$	0.040	0.052	0.050	0.046	0.041	0.041	0.054	0.052	0.048	0.041
	cvg	0.938	0.936	0.945	0.939	0.936	0.933	0.954	0.954	0.933	0.954
local1	ave	0.555	1.113	1.040	0.888	0.592	0.557	1.114	1.041	0.891	0.594
	sd	0.048	0.063	0.057	0.055	0.048	0.047	0.057	0.057	0.051	0.045
	$\widehat{\text{sd}}$	0.050	0.065	0.063	0.058	0.051	0.048	0.059	0.057	0.053	0.047
	cvg	0.938	0.942	0.953	0.944	0.938	0.945	0.952	0.946	0.947	0.941
local2	ave	0.563	1.131	1.056	0.903	0.602	0.564	1.128	1.054	0.902	0.600
	sd	0.041	0.053	0.048	0.047	0.041	0.047	0.064	0.051	0.056	0.045
	$\widehat{\text{sd}}$	0.040	0.052	0.050	0.047	0.041	0.041	0.054	0.052	0.048	0.041
	cvg	0.943	0.926	0.945	0.939	0.945	0.927	0.938	0.935	0.934	0.944
local3	ave	0.564	1.131	1.056	0.903	0.602	0.566	1.133	1.057	0.905	0.603
	sd	0.040	0.051	0.047	0.045	0.040	0.041	0.050	0.049	0.045	0.039
	$\widehat{\text{sd}}$	0.040	0.052	0.051	0.047	0.041	0.041	0.055	0.053	0.048	0.041
	cvg	0.951	0.937	0.948	0.949	0.947	0.944	0.952	0.951	0.944	0.955
MAVE	ave	0.551	1.104	1.033	0.883	0.590	0.557	1.102	1.033	0.888	0.596
	sd	0.043	0.055	0.055	0.048	0.043	0.040	0.053	0.052	0.048	0.042
PHD	ave	0.555	1.125	1.042	0.903	0.603	0.613	1.054	1.003	0.927	0.735
	sd	0.158	0.212	0.206	0.196	0.162	0.113	0.139	0.130	0.123	0.147
Semi-PHD	ave	0.551	1.104	1.033	0.883	0.590	0.575	1.099	1.033	0.893	0.609
	sd	0.054	0.069	0.068	0.061	0.054	0.049	0.067	0.064	0.058	0.053

(b) The first locally efficient estimator, local1: we obtain the estimator from the estimating equation (6) with $w_1(\mathbf{x}) = 1$.
(c) The second locally efficient estimator, local2: we obtain the estimator from the estimating equation (7) with $w_2(\mathbf{x}, \boldsymbol{\zeta})$ being a quadratic function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.
(d) The third locally efficient estimator, local3: we obtain the estimator from the estimating equation (10) with $\boldsymbol{\xi} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and w_3 is estimated through kernel regression.
(e) The refined MAVE: this is the estimator that was proposed in Xia *et al.* (2002) and is considered the currently best available estimator for central mean subspace estimation.
(f) The PHD estimator: this is the estimator that was proposed in Li (1992). It uses the first d principal eigenvectors of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ where $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$ and $\boldsymbol{\Lambda} = E[\{Y - E(Y)\}\{\mathbf{x} - E(\mathbf{x})\}\{\mathbf{x} - E(\mathbf{x})\}^{\mathrm{T}}]$ as the basis of the estimated $\mathcal{S}_{E(Y|\mathbf{x})}$.
(g) The semiparametric extension of the PHD estimator, semi-PHD: this is the estimator that was proposed in Ma and Zhu (2012). It estimates $\mathcal{S}_{E(Y|\mathbf{x})}$ through solving the sample version of the estimating equation

$$E[\{Y - E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}\{\mathbf{x}\mathbf{x}^{\mathrm{T}} - E(\mathbf{x}\mathbf{x}^{\mathrm{T}}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}] = \mathbf{0}.$$

Table 1 summarizes the true parameter values under the parameterization in Section 2 with $\mathbf{x}$ standardized, the average of the estimates the sample standard deviations of the estimates and the mean of the estimated standard deviations. We also report the empirical coverage probability of the estimates at the nominal level of 95%. The results of both scenarios for generating $\mathbf{x}$ in the first example carry almost the same messages. Obviously, all the estimates have very small biases, whereas, when the variance model contains the truth (local2 and local3), the locally efficient estimators perform competitively in comparison with the oracle estimator. The PHD estimator

Table 2. Average of estimators, ave, sample estimation standard deviation, sd, average of the estimated standard deviation, $\widehat{\mathrm{sd}}$, and 95% coverage, cvg, of various estimators in simulation 2

Method	*Statistic*	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Truth		−1	−2	−0.5	−1	−2	−2	−1.5	−1
Oracle	ave	−1.033	−2.076	−0.515	−1.038	−2.054	−2.048	−1.545	−1.023
	sd	0.107	0.138	0.081	0.076	0.143	0.164	0.107	0.095
	$\widehat{\mathrm{sd}}$	0.111	0.144	0.083	0.080	0.142	0.168	0.107	0.095
	cvg	0.956	0.930	0.939	0.942	0.949	0.942	0.936	0.944
local1	ave	−1.039	−2.090	−0.518	−1.045	−2.061	−2.052	−1.550	−1.026
	sd	0.110	0.142	0.083	0.078	0.146	0.168	0.109	0.099
	$\widehat{\mathrm{sd}}$	0.113	0.147	0.086	0.081	0.141	0.168	0.108	0.094
	cvg	0.946	0.924	0.941	0.929	0.932	0.932	0.928	0.927
local2	ave	−1.024	−2.051	−0.512	−1.026	−2.044	−2.040	−1.537	−1.020
	sd	0.112	0.152	0.083	0.087	0.158	0.178	0.117	0.112
	$\widehat{\mathrm{sd}}$	0.123	0.162	0.094	0.087	0.184	0.223	0.136	0.112
	cvg	0.949	0.940	0.948	0.941	0.941	0.946	0.936	0.928
local3	ave	−1.030	−2.072	−0.513	−1.035	−2.049	−2.041	−1.540	−1.020
	sd	0.108	0.138	0.081	0.076	0.144	0.164	0.106	0.096
	$\widehat{\mathrm{sd}}$	0.111	0.144	0.084	0.080	0.143	0.168	0.108	0.095
	cvg	0.951	0.940	0.947	0.935	0.952	0.942	0.945	0.943
MAVE	ave	−1.005	−1.883	−0.518	−1.060	−1.909	−1.859	−1.482	−1.010
	sd	0.167	0.195	0.129	0.123	0.224	0.259	0.175	0.163
PHD	ave	−1.490	−2.012	−0.974	−1.653	−2.560	−1.320	−2.404	−1.410
	sd	3.579	2.606	2.689	2.366	5.731	4.396	4.426	3.909
Semi-PHD	ave	−1.003	−1.923	−0.509	−1.029	−2.025	−1.935	−1.535	−1.044
	sd	0.166	0.190	0.117	0.146	0.362	0.344	0.270	0.295

has very large variability, whereas the variabilities of the semi-PHD estimator and the MAVE are also slightly larger than the local estimators with parametric (local2) and semiparametric (local3) models.

The conclusion of simulation 2 in Table 2 is quite different. The bias of the PHD estimator is quite substantial in this case, possibly because of the violation of the linearity condition and the constant variance condition that are required by the PHD method. The variability of the semi-PHD estimator is much larger than all the locally efficient estimators, whereas the variability of the MAVE is also generally larger than the semiparametric locally efficient estimator. Surprisingly, although, in this case, the quadratic parametric model (local2) does not capture the true variance function, the estimation variance does not deteriorate dramatically in comparison with the oracle estimator and the semiparametric model (local3) which does capture the true variance function. In fact, even when we set the weight $w_1 = 1$ (local1), its estimation variability does not increase very much. This kind of insensitivity is specific to some models and does not generalize, since the property is not shared in the previous simulation example.

Finally, we would like to point out that, from the results in both Tables 1 and 2, we see a close approximation between the sample and average estimated standard deviations; this indicates that, at sample size $n = 500$, the asymptotic properties can already be used to perform quite reliable inference in these models. This is also reflected in the 95% coverage, where the sample coverage is not too far from the nominal level.

5. An application

We further illustrate the performance of the proposed methods in analysing a data set that is related to a gender discrimination law suit of the Fifth National Bank of Springfield (Albright *et al.*, 1999). The bank was charged with paying lower salaries to its female employees than to its male employees. After removing an obvious outlier, the data set contains 207 observations. The average salary for the male employees is \$45 505 and the average for the females is \$37 262, yielding a p-value less than 10^{-6} from a two-sample t-test. However, such a naive comparison may not be suitable because there are many other social characteristics that may affect an employee's salary. We study how an employee's salary associates with his or her social characteristics.

We regard an employee's annual salary as the response variable Y. In this data set, there are four categorical and three continuous covariates. The three continuous covariates are working experience at the current bank, X_1, measured by years since an employee was hired, the employee's age, X_2, and experience at another bank before working at the Fifth National Bank, X_3, measured by the number of years at another bank. The four categorical covariates include the gender, X_4, a binary variable X_5 indicating whether the employee's job is computer related or not, a five-level categorical variable $(X_{61}, \ldots, X_{64})$ representing the employee's level of education and a six-level categorical variable denoting the employee's current job level $(X_{71}, \ldots, X_{75})$. We denote the covariate vector by $\mathbf{x} = (X_1, \ldots, X_5, X_{61}, \ldots, X_{64}, X_{71}, \ldots, X_{75})^{\mathrm{T}}$. We standardize all the continuous variables to have zero mean and unit variance in our subsequent analysis.

Because we are concerned with how the average salary changes with these social characteristics, the mean function $E(Y|\mathbf{x})$ is of our primary interest. Observing that the covariate dimension p is 14, we assume that there is a p-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5, \beta_{61}, \ldots, \beta_{64}, \beta_{71}, \ldots, \beta_{75})^{\mathrm{T}}$ (with the co-ordinates corresponding to the covariates) which satisfies

$$E(Y|\mathbf{x}) = E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}).$$

Table 3. Coefficient estimates and their standard errors for the Fifth National Bank data

Parameter	*Results for the following methods:*								
	local1		*local2*		*local3*		*MAVE estimate*	*PHD estimate*	*Semi-PHD estimate*
	Estimate	$\widehat{sd}$	*Estimate*	$\widehat{sd}$	*Estimate*	$\widehat{sd}$			
β_2	0.043	0.059	0.051	0.086	0.073	0.108	−0.110	−0.270	0.177
β_3	0.215	0.077	0.218	0.083	0.212	0.096	0.154	0.112	0.087
β_4	0.032	0.187	0.058	0.210	0.115	0.177	0.258	0.412	0.362
β_5	2.008	0.185	2.086	0.264	2.157	0.352	1.762	0.182	1.671
$\beta_{6,1}$	−0.367	0.498	−0.371	0.505	−0.330	0.404	−0.548	−0.924	−0.160
$\beta_{6,2}$	−1.068	0.527	−1.057	0.515	−0.996	0.431	−0.813	−0.512	−1.151
$\beta_{6,3}$	−0.536	0.409	−0.536	0.381	−0.478	0.309	−0.497	−0.210	−0.525
$\beta_{6,4}$	−0.959	0.575	−0.862	0.626	−0.740	0.563	−0.742	−0.425	0.299
$\beta_{7,1}$	−9.912	1.025	−10.156	1.064	−10.438	0.902	−7.211	−2.844	−9.992
$\beta_{7,2}$	−8.380	0.825	−8.631	0.931	−8.933	0.824	−6.405	−3.017	−8.776
$\beta_{7,3}$	−6.318	0.732	−6.455	0.764	−6.683	0.632	−4.736	−3.325	−7.658
$\beta_{7,4}$	−4.589	0.665	−4.707	0.671	−4.922	0.663	−2.916	−3.700	−5.841
$\beta_{7,5}$	−2.227	0.331	−2.310	0.350	−2.466	0.652	−1.123	−3.883	−3.015
$\rho_{Y,\hat{Y}}$	0.943		0.960		0.960		0.941	0.806	0.941

This is a problem of estimating the central mean subspace. Ma and Zhu (2012) demonstrated through a bootstrap procedure that a one-dimensional model is sufficient to describe the regression relationship of Y given $\mathbf{x}$ in this data set. Because an employee's salary usually depends on working experience X_1, we fix $\beta_1 = 1$ to facilitate the comparison of different inference procedures. Consequently, there are 13 free parameters in total.

We applied the three locally efficient estimators and report the estimated coefficients $\hat{\boldsymbol{\beta}}$ and the associated estimation standard errors (Table 3). For comparison, we also include the MAVE, PHD and semi-PHD estimation results. None of the three locally efficient estimates provides sufficient evidence that there is gender discrimination at the level of significance of 0.05. From the analysis, the social characteristics that positively affect an employee's salary include more working experience (both in the current position and at a previous bank), performing computer-related work, having finished some college level courses or received higher education. In addition, an employee's salary is directly linked to his or her job level. The scatter plots of the salary with respect to the estimated directions are given in Fig. 1.

Next we fit a local linear regression of Y on $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$ to obtain the fitted values $\hat{Y} = \hat{E}(Y|\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$. We then calculate the Pearson correlation coefficient $\rho_{\hat{Y},Y}$, summarized in the last row of Table 3. In terms of these correlation coefficients, the PHD method performs the worst, partly because the linearity and constant variance conditions on $\mathbf{x}$ that are required by the PHD method are violated owing to the presence of categorical covariates. The MAVE method, semi-PHD and the locally efficient estimates have comparable performance, with the parametric and semiparametric locally efficient estimators winning slightly. However, a theoretically justified inference procedure is not available for the MAVE method; hence the estimation result cannot be used to draw conclusion about the discrimination issue.

From a visual inspection of Fig. 1, we might suspect that Y could be a linear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. To examine whether this is so, we test the null hypothesis

$$H_0 : E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = a + b(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \qquad \text{for some } a \text{ and } b.$$

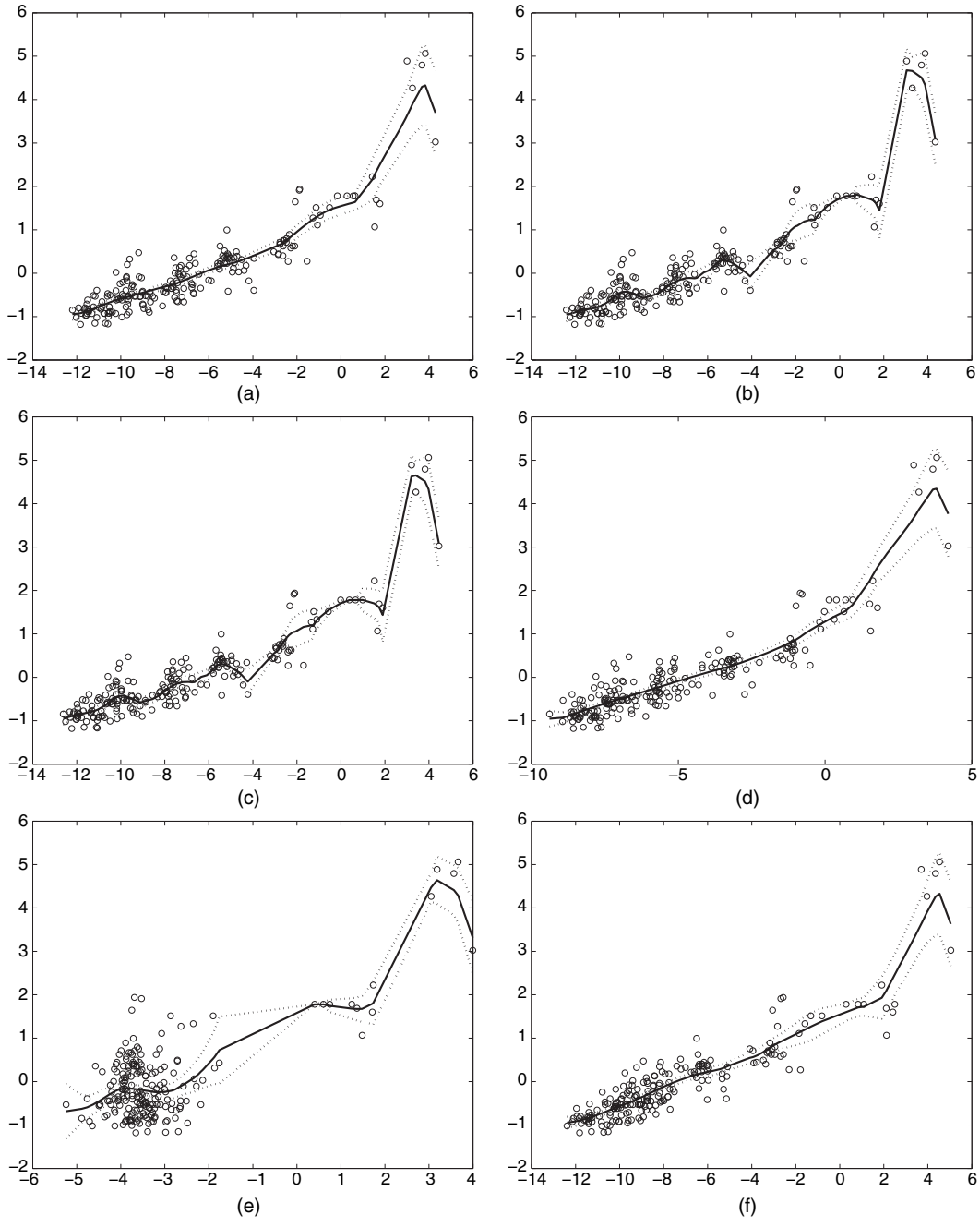

Fig. 1. Scatter plots of Y against the direction found by (a) the local1, (b) local2, (c) local3, (d) MAVE, (e) PHD and (f) semi-PHD methods: ———, local linear approximation; ·····, associated 95% pointwise confidence interval

Following Fan *et al.* (2001), we define a generalized likelihood ratio test statistic

$$\lambda_n = \frac{n}{2}\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1},$$

where

$$\mathrm{RRS}_0 = \sum_{i=1}^{n}\{Y_i - \hat{a} - \hat{b}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}^2,$$

$$\mathrm{RRS}_1 = \sum_{i=1}^{n}\{Y_i - \hat{m}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_i)\}^2.$$

Here, $\hat{a}$ and $\hat{b}$ are estimated from ordinary least squares, $\hat{m}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$ is a local linear estimate of the mean function $E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\hat{\boldsymbol{\beta}}$ is an arbitrary root-n-consistent estimator of $\boldsymbol{\beta}$. For all the proposed locally efficient estimators of $\boldsymbol{\beta}$, we obtain that the p-values are all less than 10^{-4}. This indicates that a linear model is not sufficient to capture fully the regression information of Y onto $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$.

6. Discussion

Throughout the paper, we have not assumed the usual linearity condition or the constant variance condition. We point out here that, if the linearity condition holds, the efficiency bound does not change for central mean subspace estimation. However, the linearity condition will contribute to some computational simplification when we choose to set $w(\mathbf{x})$ to be a function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. This is because we can simply plug $E(\mathbf{x}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}\mathbf{x}$ into the estimation equation instead of estimating it non-parametrically. However, as soon as other choices are taken for $w(\mathbf{x})$, the linearity condition has no particular use. In contrast, the constant variance condition does not contribute to the efficiency bound or to the computational simplicity. It is therefore a redundant condition in the context of efficient estimation.

Throughout we assume that the structural dimension d of $\mathcal{S}_{E(Y|\mathbf{x})}$ is known *a priori*. How to estimate the structural dimension when it is unknown has been extensively studied in the literature. See for example the sequential test procedure in Cook and Li (2004), the cross-validation procedure in Xia *et al.* (2002) and Xia (2007), the Bayesian information criterion BIC type of criterion in Zhu *et al.* (2006) and the bootstrap procedure in Ye and Weiss (2003), Li and Dong (2009), Dong and Li (2010), Ma and Zhu (2012), etc. Thus we do not elaborate this issue further in this paper.

Various model extensions have been considered in the dimension reduction literature. For example, in the partial dimension reduction problems (Li *et al.*, 2003), it is assumed that $E(Y|\mathbf{x}) = E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1, \mathbf{x}_2)$. Here, $\mathbf{x}_1$ is a covariate subvector of $\mathbf{x}$ that the dimension reduction procedure focuses on, whereas $\mathbf{x}_2$ is a covariate subvector that is known on the basis of scientific understanding or convention to enter the model directly. We can see that the semiparametric analysis and the efficient estimation results that are derived here can be adapted to these models, through changing $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ to $(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1, \mathbf{x}_2)$ in all the corresponding functions and expectations while everything else remains unchanged. Another extension is groupwise dimension reduction (Li *et al.*, 2010), where the model $E(Y|\mathbf{x}) = \Sigma_{i=1}^{k} m_i(Y, \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}_i)$ is considered. The semiparametric analysis in such models requires separate investigation and it will be interesting to study efficient estimation in these problems.

Acknowledgements

Ma's work was supported by the National Science Foundation (grants DMS-1000354 and

DMS-1206693) and the National Institute of Neurological Disorders and Stroke (grant R01-NS073671). Liping Zhu is also affiliated with the Key Laboratory of Mathematical Economics, Ministry of Education. Zhu's work is supported by the National Natural Science Foundation of China (grants 11071077 and 11371236), the innovation programme of the Shanghai Municipal Education Commission (grant 13ZZ055), the Pujiang project of the Science and Technology Commission of Shanghai Municipality (grant 12PJ1403200) and the 'Program for new century excellent talents', Ministry of Education of China (grant NCET-12-0901).

Appendix A

A.1. Derivation of $\mathbf{S}_{\text{eff}}$ *in model (1)*

From Ma and Zhu (2012), we have

$$\begin{aligned}\Lambda^{\perp} &= ([\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}]\varepsilon : \forall \boldsymbol{\alpha}(\mathbf{x})) \\ &= (\{Y - E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\}[\boldsymbol{\alpha}(\mathbf{x}) - E\{\boldsymbol{\alpha}(\mathbf{x})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}] : \forall \boldsymbol{\alpha}(\mathbf{x}))\end{aligned}$$

and $\Lambda = \Lambda_1 \oplus \Lambda_2 + \Lambda_m$, where

$$\begin{aligned}\Lambda_1 &= \{f(\mathbf{x}) : \forall f \text{ subject to } F(f) = 0\}, \\ \Lambda_2 &= \{f(\varepsilon, \mathbf{x}) : \forall f \text{ subject to } E(f|\mathbf{x}) = 0, E(\varepsilon f|\mathbf{x}) = 0\}, \\ \Lambda_m &= \left\{ \frac{\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})}{\eta_2(\varepsilon, \mathbf{x})} h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) : \forall h \right\}.\end{aligned}$$

Here $\varepsilon = Y - m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})$ is the derivative of η_2 with respect to ε. We calculate the residual after projecting a function in Λ_m to Λ_2 to obtain

$$\Lambda'_m = \Pi(\Lambda_m|\Lambda_2^{\perp}) = \left\{ \frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) : \forall h \right\}.$$

Thus, $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda'_m$.

Straightforward calculation yields

$$\mathbf{S}_{\beta} = -\operatorname{vecl}\left\{ \frac{\mathbf{x}\,\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})\,\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\eta_2(\varepsilon, \mathbf{x})\,\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})} \right\}.$$

Since

$$\begin{aligned}\mathbf{S}_{\beta} = &-\operatorname{vecl}\left[\left\{ \frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} + \frac{\eta'_{2\varepsilon}(\varepsilon, \mathbf{x})}{\eta_2(\varepsilon, \mathbf{x})} \right\} \frac{\mathbf{x}\,\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})} \right] + \operatorname{vecl}\left[\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} \frac{E\{\mathbf{x}E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}{E\{E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}} \frac{\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})} \right] \\ &+ \operatorname{vecl}\left(\frac{\varepsilon}{E(\varepsilon^2|\mathbf{x})} \left[\mathbf{x} - \frac{E\{\mathbf{x}E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}}{E\{E(\varepsilon^2|\mathbf{x})^{-1}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}} \right] \frac{\partial m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{\partial(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})} \right),\end{aligned}$$

and we can easily verify that the first component is in Λ_2, the second component is in Λ'_m and the third component is in $\Lambda^{\perp}$; hence we obtain $\mathbf{S}_{\text{eff}}$.

A.2. List of regularity conditions

Condition 1. The univariate kernel function $K(\cdot)$ is symmetric, has compact support and is Lipschitz continuous on its support. It satisfies

$$\begin{aligned}&\int K(u)\,\mathrm{d}u = 1, \\ &\int u^i K(u)\,\mathrm{d}u = 0, \qquad 1 \leqslant i \leqslant m-1, \\ &0 \neq \int u^m K(u)\,\mathrm{d}u < \infty.\end{aligned}$$

Condition 2. The probability density functions of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}$, which are denoted respectively by $f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $f_2(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$, and the variance function $E(\varepsilon^2|\mathbf{x})$ are bounded away from 0 and ∞.

Condition 3. Let

$$\mathbf{r}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = E\{\mathbf{a}_1(\mathbf{x}, Y)|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}\, f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$$

and

$$\mathbf{r}_2(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}) = E(\varepsilon^2|\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})\, f_2(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}),$$

where $\mathbf{a}_1(\mathbf{x}, Y)$ can be any of Y, $w_3(\boldsymbol{\xi})$ or $\mathbf{x}\, w_3(\boldsymbol{\xi})$. The $(m-1)$th derivatives of $\mathbf{r}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, $\mathbf{r}_2(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$, $f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, $f_2(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})$ and $m(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ are locally Lipschitz continuous.

Condition 4. $E(\|\mathbf{x}\|^4) < \infty$, $E(Y^4) < \infty$ and $E\{\|\mathbf{m}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\|^4\} < \infty$.

Condition 5. The bandwidths satisfy $nh_k^{2m}h_l^{2m} \to 0$ and $nh_1^{2(m-1)}h_l^{2m} \to 0$ and $nh_k^d h_l^d \to \infty$ for $1 \leqslant k < l \leqslant 4$.

References

Albright, S. C., Winston, W. L. and Zappe, C. J. (1999) *Data Analysis and Decision Making with Microsoft Excel*. Pacific Grove: Duxbury.

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

Cook, R. D. and Li, B. (2002) Dimension reduction for conditional mean in regression. *Ann. Statist.*, **30**, 455–474.

Cook, R. D. and Li, B. (2004) Determining the dimension of iterative Hessian transformation. *Ann. Statist.*, **32**, 2501–2531.

Cook, R. D. and Ni, L. (2005) Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Statist. Ass.*, **100**, 410–428.

Cook, R. D. and Ni, L. (2006) Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, **93**, 65–74.

Delecroix, M., Härdle, W. and Hristache, M. (2003) Efficient estimation in conditional single-index regression. *J. Multiv. Anal.*, **86**, 213–226.

Dong, Y. and Li, B. (2010) Dimension reduction for non-elliptically distributed predictors: second-order moments. *Biometrika*, **97**, 279–294.

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

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

Horowitz, J. and Härdle, W. (1996) Direct semiparametric estimation of single-index models with discrete covariates. *J. Am. Statist. Ass.*, **91**, 1632–1640.

Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.*, **86**, 316–342.

Li, K. C. (1992) On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Ass.*, **87**, 1025–1039.

Li, B., Cook, R. D. and Chiaromonte, F. (2003) Dimension reduction for conditional mean in regression with categorical predictors. *Ann. Statist.*, **31**, 1636–1668.

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

Li, K. C. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

Li, L., Li, B. and Zhu, L. X. (2010) Groupwise dimension reduction. *J. Am. Statist. Ass.*, **105**, 1188–1201.

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. P. (2012) A semiparametric approach to dimension reduction. *J. Am. Statist. Ass.*, **107**, 168–179.

Ma, Y. and Zhu, L. P. (2013) Efficiency loss caused by linearity condition in dimension reduction. *Biometrika*, **100**, 371–383.

Maity, A., Ma, Y. and Carroll, R. J. (2007) Efficient estimation of population-level summaries in general semiparametric regression models. *J. Am. Statist. Ass.*, **102**, 123–139.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

Van Keilegom, I. and Carroll, R. J. (2007) Backfitting versus profiling in general criterion functions. *Statist. Sin.*, **17**, 797–816.

Xia, Y. (2007) A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, **35**, 2654–2690.

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

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.
Ye, Z. and Weiss, R. E. (2003) Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Ass.*, **98**, 968–979.
Yoo, P. and Cook, R. D. (2007) Optimal sufficient dimension reduction for the conditional mean in multivariate regressions. *Biometrika*, **94**, 231–242.
Zhu, L. X., Miao, B. Q. and Peng, H. (2006) On sliced inverse regression with large dimensional covariates. *J. Am. Statist. Ass.*, **101**, 630–643.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplement for "On estimation efficiency of the central mean subspace"'.