

On efficient dimension reduction with respect to a statistical functional of interest

WEI LUO, BING LI, AND XIANGRONG YIN

Abstract

We introduce a new sufficient dimension reduction framework that targets a statistical functional of interest, and propose an efficient estimator for the semiparametric estimation problems of this type. The statistical functional covers a wide range of applications, such as conditional mean, conditional variance, and conditional quantile. We derive the general forms of the efficient score and efficient information as well as their specific forms for three important statistical functionals: the linear functional, the composite linear functional, and the implicit functional. In conjunction with our theoretical analysis we also propose a class of one-step Newton-Raphson estimators and show by simulations that they substantially outperform existing methods. Finally, we apply the new method to construct the central mean and central variance subspaces for a data set involving the physical measurements and age of abalones, which exhibits a strong pattern of heteroscedasticity.

Key words and phrases. Central subspace; Conditional mean, variance, and quantile; Efficient information; Efficient score; Fréchet derivative and its representation; Projection; Tangent space.

1 Introduction

The purpose of this paper is twofold: to introduce a new framework for sufficient dimension reduction that targets a statistical functional of interest, and to develop semiparametrically efficient estimators for problems of this type.

Let X be a p -dimensional random vector and Y be a random variable. In classical sufficient dimension reduction (SDR) we are interested in a lower dimensional linear predictor $\zeta^\top X$, where ζ is a $p \times d$ matrix with $d < p$, such that Y is independent of X given $\zeta^\top X$. That is, the conditional distribution of Y given X is the same as that of Y given $\zeta^\top X$. In this problem, the identifiable parameter is $\text{span}(\zeta)$, the subspace of \mathbb{R}^p spanned by the columns of ζ . Under mild conditions there exists a unique smallest subspace that satisfies this condition, and it is called the *central subspace*. See Li (1991, 1992), Cook and Weisberg (1991), Cook (1994, 1996, 1998). For a general discussion of the sufficient conditions for the central subspace to exist, see Yin, Li, and Cook (2008). The SDR provides us a mechanism to reduce the dimension of the predictor while preserving the conditional distribution of Y given X .

In many applications our interests are in some specific aspects of the conditional distribution $P_{Y|X}$. For example, in nonparametric regression we are interested in the conditional mean $E(Y|X)$; in median regression we are interested in the conditional median $M(Y|X)$, and in volatility analysis we are interested in the conditional variance $\text{var}(Y|X)$, and in supervised classification we are interested in the class label of Y given its covariates X . To illuminate this point further, let us consider the model

$$Y = \mu(X_1) + \sigma(X_1 + X_2)\varepsilon, \quad (1)$$

where μ and σ are unknown functions. In this case the central subspace, which is the 2-dimensional subspace of \mathbb{R}^p spanned by $(1, 0, \dots, 0)$ and $(0, 1, 0, \dots, 0)$, only tells us the sufficient predictors can be any linear combination of X_1 and X_2 , but it does not tell us that the conditional mean is a function of X_1 , and the conditional variance is a function of $X_1 + X_2$. Thus, the information provided by the central subspace is clearly inadequate if we want to build a model like (1).

Under these circumstances it makes sense to reformulate sufficient dimension reduction to target a specific functional, so as to provide a more nuanced picture of the relation between X and Y than offered by the central subspace. Several such efforts have been made over the past decade or so. For example, Cook and Li (2002) introduced the central mean subspace, which is defined by the relation $E(Y|X) = E(Y|\alpha^\top X)$, where α is minimal in the same sense as it is in the central subspace. Yin and Cook (2002) introduced the k th central moment subspace through the relation $E(Y^k|X) = E(Y^k|\alpha^\top X)$. Zhu and Zhu (2009) introduced the central variance subspace by requiring that $\text{var}(Y|X)$ is a function of $\alpha^\top X$. Zhu, Dong, and Li (2012) introduced a general class of estimating equations for single-index conditional variance. The Minimum Average Variance Estimator (MAVE) by Xia, Tong, Li, and Zhu (2002) also targets the central mean subspace. Kong and Xia (2012) introduced an adaptive quantile estimator for single-index quantile regression, which targets the conditional quantile. It turns out the space spanned by the columns of α in the above relations are all subspaces of the central subspace. They provide refined structures for the central subspace. As a consequence, $\alpha^\top X$ can be written as $\beta^\top \zeta^\top X$, and we can refine central subspace based on the sufficient predictor $\zeta^\top X$. For convenience, we reset $\zeta^\top X$ as X , and use \tilde{X} to represent the original predictor throughout the rest of the paper.

The first goal of this paper is to unify these problems by introducing a general dimension reduction paradigm with respect to statistical functional T of the conditional density of Y given X , say $\eta(x, y)$, through the following statement

$$T(\eta(x, \cdot)) \text{ is a function of } \beta^\top x. \quad (2)$$

Note that sufficient dimension reduction for conditional mean, conditional moments, and conditional variance discussed in the last paragraph are all special cases

of relation (2). The minimal subspace $\text{span}(\beta)$ of \mathbb{R}^p that satisfies this relation is called the T -central subspace.

The second, and the main goal of this paper is to develop semiparametrically efficient estimators for the T -central subspace. In a series of recent papers, Ma and Zhu (2012, 2013a, 2013b) use semiparametric theory to study sufficient dimension reduction and develop semiparametrically efficient estimators of the central subspace. These are related to the earlier developments by Li and Dong (2009) and Dong and Li (2010), which use estimating equations to relax the elliptical distribution assumption for sufficient dimension reduction. We extend Ma and Zhu's approach to find semiparametrically efficient estimator for the T -central subspace. We derive the general formulas for the efficient score and efficient information for the semiparametric family specified by the relation (2), and further deduce their specific forms for three important statistical functionals: the linear functionals (L-functionals), the composite linear functionals (C-functionals), and implicit functionals (I-functionals). These functionals cover a wide range of applications. For example, all conditional moments are L-functionals, all conditional cumulants (see, for example, McCullagh, 1987) are C-functionals, and quantities such as conditional median, conditional quantile, and conditional support vector machine (Li, Artemiou, and Li, 2011) are I-functionals.

Using the formulas for efficient score and efficient information, we propose a one-step Newton-Raphson algorithm to implement semiparametrically efficient estimation. Compared with the semiparametric estimators of Ma and Zhu (2013a), our algorithm has two distinct and attractive features. First, since our algorithm relies on the MAVE-type procedure for minimization, it can be implemented by iterations of a least squares problem without resorting to high-dimensional search-based optimization. Second, unlike Ma and Zhu (2013a), our method does not require any specific parameterization of β that potentially restricts the generality of their method.

The rest of the paper is organized as follows. In Section 2 we give a general formulation of sufficient dimension reduction with respect to a statistical functional of interest. To set the stage for further development, we lay out the semiparametric structure of our problem in Section 3. In Section 4 we derive the efficient score and efficient information for a general statistical functional. In Sections 5, 6, and 7 we further deduce the specific forms of the efficient score and efficient information for the L-, C-, and I-functionals. In section 8 we discuss the effect of estimating the central subspace on the efficient score. In section 9 we develop the one-step Newton-Raphson estimation procedure for semiparametrically efficient estimation. In section 10 we conduct simulation studies to compare our method with other methods, and in Section 11 we apply our method to a data set. Some concluding remarks are made in Section 12. The proofs of some technical results are given in

an External Appendix.

The following notations will be consistently used throughout the rest of the paper. The symbol I_k denotes the $k \times k$ dimensional identity matrix; e_k denotes a vector whose k th entry is 1 and other entries are 0; $\perp\!\!\!\perp$ indicates independence or conditional independence between two random elements — that is, $A \perp\!\!\!\perp B$ means A and B are independent, and $A \perp\!\!\!\perp B|C$ means A and B are independent given C . For integers s and t , \mathbb{R}^s denotes the s dimensional Euclidean space, and $\mathbb{R}^{s \times t}$ denotes the set of $s \times t$ dimensional matrices. For a function with multiple arguments, say $f(x, y, z)$, we use the dot notation to represent mappings of a subset of the arguments. For example, $f(\cdot, y, z)$ represents the mapping $x \mapsto f(x, y, z)$ where y and z are fixed, and $f(\cdot, \cdot, z)$ represents the mapping $(x, y) \mapsto f(x, y, z)$ where z is fixed. We use superscripts of X to index components and subscripts of X to index subjects. Thus X_i^j means the j th component of the i th observation in a sample X_1, \dots, X_n . However, a^i represents power when a is not X .

2 Dimension reduction for conditional statistical functional

Let $(\Omega_X, \mathcal{B}_X, \mu_X)$ and $(\Omega_Y, \mathcal{B}_Y, \mu_Y)$ be σ -finite measure spaces, where $\Omega_X \subseteq \mathbb{R}^d$ and $\Omega_Y \subseteq \mathbb{R}$ and \mathcal{B}_X and \mathcal{B}_Y are σ -fields of Borel sets in Ω_X and Ω_Y . Let (X, Y) be a pair of random elements that takes values in $(\Omega_X \times \Omega_Y, \mathcal{B}_X \times \mathcal{B}_Y)$. Let \mathcal{M} be a family of densities of (X, Y) with respect to $\mu = \mu_X \times \mu_Y$. We assume that \mathcal{M} is a *semiparametric* family; that is, there exist $\Theta \subseteq \mathbb{R}^r$ and a family \mathcal{F} of functions $\phi : \Omega_X \times \Omega_Y \times \Theta \rightarrow \mathbb{R}$ such that

$$\mathcal{M} = \cup\{\mathcal{M}_\theta : \theta \in \Theta\}, \text{ where } \mathcal{M}_\theta = \{\phi(\cdot, \cdot, \theta) : \phi \in \mathcal{F}\}.$$

Furthermore, we assume that each $\phi \in \mathcal{F}$ can be factorized into $\lambda(x)\eta(x, y, \theta)$ where λ is the marginal density of X , $\eta(x, \cdot, \theta)$ is the conditional density of Y given X . The real assumption in this factorization is that θ appears only in the conditional density.

As an illustration, consider the single index model where

$$Y = m(\beta^\top X) + \varepsilon, \quad \beta \in \mathbb{R}^d, \quad X \perp\!\!\!\perp \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad X \sim N(0, \Sigma),$$

and m is an unknown function. Since m is unknown, β is identified up to a proportional constant. To avoid the trivial case let us assume it has at least one nonzero component, and further assume this is the first component for convenience. We can then assume without loss of generality $\beta^\top = (1, \theta^\top)$ where $\theta \in \mathbb{R}^{d-1}$. Then

$$\beta = e_1 + \Gamma\theta, \quad m(\beta^\top X) = m(X^1 + \theta^\top \Gamma^\top X),$$

where Γ^\top is the $d \times (d-1)$ matrix $(0, I_{d-1})$. In this case, $\lambda(x)$ is the p.d.f. of $N(0, \Sigma)$ and $\eta(x, \cdot, \theta)$ is the p.d.f. of $N(m(x^1 + \theta^\top \Gamma^\top x), \sigma^2)$ for a given x .

Now let \mathcal{F}_1 denote the family $\{\eta : \phi \in \mathcal{F}\}$, and

$$\mathcal{L} = \{\lambda : \phi \in \mathcal{F}\}, \quad \mathcal{H}_\theta = \{\eta(\cdot, \cdot, \theta) : \eta \in \mathcal{F}_1\}, \quad \mathcal{H} = \cup\{\mathcal{H}_\theta : \theta \in \Theta\}.$$

We assume that \mathcal{M} contains the true density of (X, Y) . That is, there exist $\theta_0 \in \Theta$, $\lambda_0 \in \mathcal{L}$, and $\eta_0 \in \mathcal{F}_1$ such that $\phi_0(x, y, \theta_0) = \lambda_0(x)\eta_0(x, y, \theta_0)$ is the true density of (X, Y) . For convenience, we abbreviate $\phi_0(x, y, \theta_0)$, $\eta_0(x, y, \theta_0)$, and \mathcal{M}_{θ_0} as $\phi_0(x, y)$, $\eta_0(x, y)$, and \mathcal{M}_0 .

Let \mathcal{G} be a class of densities of Y with respect to μ_Y that contains all $\eta(x, \cdot, \theta)$ for $\eta \in \mathcal{F}_1$, $\theta \in \Theta$, and $x \in \Omega_X$. Let $T : \mathcal{G} \rightarrow \mathbb{R}$ be a mapping from \mathcal{G} to \mathbb{R} . Such mappings are called statistical functionals. The functional T induces the random variable

$$x \mapsto T(\eta(x, \cdot, \theta)),$$

on Ω_X , which we write as $T(\eta(X, \cdot, \theta))$. Following the convention of the conditional expectation, we write $T(\eta_0(X, \cdot, \theta_0))$ as $T(Y|X)$. This random variable can be used to characterize a wide variety of features of a conditional density $\eta(x, \cdot, \theta)$ that might interest us, as detailed by the following example.

Example 1 Let $T : \mathcal{G} \rightarrow \mathbb{R}$ be the functional $g \mapsto \int_{\Omega_Y} yg(y)d\mu_Y$. Then, each $\eta(\cdot, \cdot, \theta) \in \mathcal{H}$ uniquely defines the mapping

$$x \mapsto T(\eta(x, \cdot, \theta)) = \int_{\Omega_Y} y\eta(x, y, \theta)d\mu_Y(y).$$

That is, $T(\eta(X, \cdot, \theta))$ is the conditional expectation $E(Y|X)$ under $\eta(\cdot, \cdot, \theta)$.

Let $T : \mathcal{G} \rightarrow \mathbb{R}$ be the mapping

$$T(g) = \int y^2g(y)d\mu_Y(y) - (\int yg(y)d\mu_Y(y))^2.$$

Then, $T(\eta(X, \cdot, \theta))$ is the conditional variance $\text{var}(Y|X)$ under the conditional density $x \mapsto \eta(x, \cdot, \theta)$.

Let $T : \mathcal{G} \rightarrow \mathbb{R}$ be the functional defined by the equation in m

$$\int \text{sgn}(y - m)g(y)d\mu_Y(y) = 0, \tag{3}$$

where $\text{sgn}(a)$ is the sign function that takes the value 1 if $a \geq 0$ and takes the value -1 if $a < 0$. The solution to (3) is the median of Y . Each $\eta(\cdot, \cdot, \theta) \in \mathcal{H}$ uniquely defines the mapping $T(\eta(X, \cdot, \theta))$, which is the conditional median of Y given X under the conditional density $x \mapsto \eta(x, \cdot, \theta)$. \square

We now give a rigorous definition of the T -central subspace.

Definition 1 *If there is a matrix $\gamma \in \mathbb{R}^{d \times u}$, with $u < d$, such that $T(\eta_0(X, \cdot, \theta_0))$ is measurable with respect to $\sigma(\gamma^\top X)$, then we call $\text{span}(\gamma)$ a sufficient dimension reduction subspace for T . The intersection of all such spaces is called the central subspace for conditional functional T , or the T -central subspace.*

We denote the T -central subspace by $\mathcal{S}_{T(Y|X)}$. For example, if T is the conditional mean functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central mean subspace, which we write as $\mathcal{S}_{E(Y|X)}$; if T is the conditional median functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central median subspace, which we write as $\mathcal{S}_{M(Y|X)}$; if T is the conditional variance functional, then $\mathcal{S}_{T(Y|X)}$ becomes the central variance subspace, which we write as $\mathcal{S}_{V(Y|X)}$. It is easy to see that $\mathcal{S}_{T(Y|X)} \subseteq \mathcal{S}_{Y|X}$: this is because $Y \perp\!\!\!\perp X | \beta^\top X$ implies

$$T(Y|X) = E[T(Y|X)|X] = E[T(Y|X)|X, \beta^\top X] = E[T(Y|X)|\beta^\top X].$$

In the following, we denote the dimension of the $\mathcal{S}_{T(Y|X)}$ by s and any basis matrix of $\mathcal{S}_{T(Y|X)}$ (of dimension $d \times s$) as β .

3 Formulation of the semiparametric problem

To set the stage for further development we first outline the basic semiparametric structure of our problem. Let $L_2(\phi_0) = \{r : \int r^2 \phi_0 d\mu < \infty\}$. Let $\langle \cdot, \cdot \rangle_{\phi_0}$ and $\|\cdot\|_{\phi_0}$ denote the inner product and norm in $L_2(\phi_0)$. For a technical reason, it is easier to work with an embedding of \mathcal{M} into $L_2(\phi_0)$, defined as

$$R : \phi \mapsto 2(\phi^{1/2} - \phi_0^{1/2})/\phi_0^{1/2} \equiv r.$$

Let $R(\mathcal{M}) = \{R(\phi) : \phi \in \mathcal{M}\}$. This transformation ensures that $R(\mathcal{M}) \subseteq L_2(\phi_0)$; whereas additional assumptions are needed to ensure $\mathcal{M} \subseteq L_2(\phi_0)$. Also note that $R(\phi_0)$ is the 0 element in $L_2(\phi_0)$.

A *curve* in $R(\mathcal{M}_0)$ that passes through $r_0 = 0$ is any mapping $\alpha \mapsto r_\alpha(\cdot)$ from $[0, 1) \rightarrow R(\mathcal{M}_0)$ that is Fréchet differentiable at $\alpha = 0$. That is, there is a member \dot{r}_0 of $L_2(\phi_0)$ such that

$$\|r_\alpha - r_0 - \dot{r}_0 \alpha\|_{\phi_0} = o(|\alpha|).$$

The *tangent space* \mathcal{T}_ϕ of $R(\mathcal{M}_0)$ at r_0 is the closure of the subspace of $L_2(\phi_0)$ spanned by \dot{r}_0 along all curves.

Let $\dot{r}_0 \in [L_2(\phi_0)]^r$ be the score with respect to θ ; that is,

$$\|R(\phi_0(\cdot, \theta)) - R(\phi_0(\cdot, \theta_0)) - \dot{r}_0^\top (\theta - \theta_0)\|_{\phi_0} = o(\|\theta - \theta_0\|).$$

Let $\Pi(\mathring{r}_0 | \mathcal{T}_\phi^\perp)$ be the componentwise projection of the random vector \mathring{r}_0 on to the orthogonal complement of the tangent space \mathcal{T}_ϕ . This projection is called the efficient score, and we denote it by $S_{\text{eff}}(X, Y, \theta_0)$. The matrix

$$J_{\text{eff}}(\theta_0) = E[S_{\text{eff}}(X, Y, \theta_0)S_{\text{eff}}^\top(X, Y, \theta_0)]$$

is called the efficient information. Now let $(X_1, Y_1), \dots, (X_n, Y_n)$ be an i.i.d. sample of (X, Y) . For a function h of (x, y) , let $E_n h(X, Y)$ denote the sample average of $h(X_1, Y_1), \dots, h(X_n, Y_n)$. Under some conditions, if $\hat{\theta}$ is the solution to the estimating equation

$$E_n S_{\text{eff}}(X, Y, \theta) = 0, \quad (4)$$

then $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, J_{\text{eff}}^{-1}(\theta_0))$. Moreover, for any estimate $\tilde{\theta}$ of θ_0 that is regular with respect to \mathcal{T}_ϕ , $\sqrt{n}(\tilde{\theta} - \theta_0)$ can be decomposed as $\sqrt{n}(\hat{\theta} - \theta_0) + \Delta_n$ where two terms are asymptotically independent. This result, well known in the semiparametric literature as the Hájek-LeCam convolution theorem, implies $\hat{\theta}$ has the smallest asymptotic variance among all regular estimators with respect to \mathcal{T}_ϕ . That is, $\hat{\theta}$ is semiparametrically efficient. For a comprehensive exposition of this theory, see Bickel, Klaassen, Ritov, and Wellner (1993, Chapter 3) or van der Vaart (1998, Chapter 25).

We now investigate how the sufficient dimension reduction in Definition 1 specifies the semiparametric family \mathcal{M} , and what is the meaning of the parameter θ in this context. Since our goal is to estimate $\text{span}(\beta)$, we need fewer parameters than ds . In fact, the set $\{\text{span}(\beta) : \beta \in \mathbb{R}^{d \times s}, \text{rank}(\beta) = s\}$ is a Grassmann manifold, which has dimension $s(d-s)$ (see, for example, Edelman, Arias, and Smith, 1998). There always exists a smooth parameterization $\beta = \beta(\theta)$, where $\theta \in \mathbb{R}^{s(d-s)}$, because $\text{span}(\beta)$ is determined if a certain $s \times s$ submatrix of β is fixed as I_s and the complementary $(d-s) \times s$ block has free varying entries. The specific form of the parameterization is not important to us.

Let $\sigma_\theta(X)$ be the σ -field generated by $\beta^\top(\theta)X$. Because $T(\eta(X, \cdot, \theta))$ is measurable with respect to $\sigma_\theta(X)$ if and only if

$$T(\eta(X, \cdot, \theta)) = E[T(\eta(X, \cdot, \theta)) | \sigma_\theta(X)],$$

the semiparametric family for our purpose is $\mathcal{M} = \cup\{\mathcal{M}_\theta : \theta \in \mathbb{R}^{s(d-s)}\}$, where

$$\mathcal{M}_\theta = \{\phi(\cdot, \cdot, \theta) : \phi \in \mathcal{F}, T(\eta(x, \cdot, \theta)) = E_x[T(\eta(X, \cdot, \theta)) | \sigma_\theta(X)]_x \quad \forall x \in \Omega_X\}.$$

Here, for a sub- σ -field \mathcal{A} of \mathcal{B}_X and a function $f(X, Y)$, $E[f(X, Y) | \mathcal{A}]_x$ denotes the evaluation of the conditional expectation $E[f(X, Y) | \mathcal{A}]$ at x .

In this paper we will focus on the development of the efficient score, the efficient information, and an accompanying estimation procedure, but will not give a

rigorous proof of the asymptotic results (including the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ and the convolution theorem), because it would far exceed the scope of the paper and because we do not expect the proof will fundamentally deviate from that given in Ma and Zhu (2013a). In addition, as mentioned earlier, by design our method is applied to the sufficient predictor corresponding to the central subspace, whose dimension is relatively low. Consequently, we expect no surprises as regards the validity of \sqrt{n} -rate of convergence of our estimator. In the meantime, our simulation studies provide strong evidence that the our efficient estimator does approach the theoretical semiparametric variance bound for modestly large sample sizes.

4 Efficient score and efficient information

In this section we derive the efficient score and efficient information for the semi-parametric problem set up in section 3. To this end we first derive the tangent space \mathcal{T}_ϕ for a fixed $\theta \in \Theta$. Let \mathcal{T}_η be the tangent space of $R(\mathcal{H}_\theta)$ at $R(\eta_0(\cdot, \cdot, \theta)) = 0$, and \mathcal{T}_λ be the tangent space of $R(\mathcal{L})$ at $R(\lambda_0) = 0$.

Proposition 1 *The following relations hold:*

1. $\mathcal{T}_\phi = \mathcal{T}_\eta + \mathcal{T}_\lambda$;
2. $\mathcal{T}_\eta \perp \mathcal{T}_\lambda$ in terms of the inner product in $L_2(\phi_0)$;
3. $\mathcal{T}_\phi^\perp = \mathcal{T}_\lambda^\perp \cap \mathcal{T}_\eta^\perp$.

This proposition was verified and used in Ma and Zhu (2012, 2013a). Since the family \mathcal{L} has no constraint, its tangent space is straightforward, as given in the next proposition, which is taken from Bickel et al (1993, page 52).

Proposition 2 \mathcal{T}_λ consists of all functions h in $L_2(\lambda_0)$ with $E_{\lambda_0} h(X) = 0$.

To compute \mathcal{T}_η , we introduce a new functional for each fixed $x \in \Omega_X$. Let $\mathcal{H}_{\theta,x}$ be the class of densities $\{\eta(x, \cdot, \theta) : \eta \in \mathcal{H}_\theta\}$. Let R_x denote the mapping

$$\begin{aligned} R_x : \mathcal{H}_{\theta,x} &\rightarrow L_2(\eta_0(x, \cdot, \theta)), \\ \eta(x, \cdot, \theta) &\mapsto 2[\eta^{1/2}(x, \cdot) - \eta_0^{1/2}(x, \cdot, \theta)]/\eta_0^{1/2}(x, \cdot, \theta). \end{aligned} \quad (5)$$

This mapping is different from R , which is from \mathcal{M} to $L_2(\phi_0)$. Nevertheless, note that $R(\eta(\cdot, \cdot, \theta))(x, y) = R_x(\eta(x, \cdot, \theta))(y)$. Let

$$T_x : L_2(\eta_0(x, \cdot, \theta)) \rightarrow \mathbb{R}, \quad r(x, \cdot, \theta) \mapsto T \circ R_x^{-1}(r(x, \cdot, \theta)). \quad (6)$$

The Fréchet derivative of T_x at $r(x, \cdot, \theta)$ is denoted by $\dot{T}_x(r(x, \cdot, \theta))$. This is a bounded linear functional from $L_2((\eta_0(x, \cdot, \theta)))$ to \mathbb{R} .

Theorem 1 Suppose, for each $x \in \Omega_X$, the functional T_x is Fréchet differentiable at θ . Let $\tau(x, \cdot, \theta)$ be the Riesz representation of $\dot{T}_x(0)$ and assume $\tau(\cdot, \cdot, \theta) \in L_2(\phi_0)$. Then

$$\mathcal{T}_\eta \subseteq \{[h(x) - E(h(X)|\sigma_\theta(X))_x]\tau(x, y, \theta) + g(x) : h, g \in L_2(\lambda_0)\}^\perp \equiv \mathcal{U}^\perp. \quad (7)$$

Moreover, if, for each $x \in \Omega_X$, the function $r(x, \cdot, \theta) \mapsto T_x(r(x, \cdot, \theta))$ is continuously Fréchet differentiable in a neighborhood of $0 \in L_2(\eta_0(x, \cdot, \theta))$, then $\mathcal{T}_\eta \supseteq \mathcal{U}^\perp$.

The proof of this theorem is technical and is presented in the External Appendix (Section I). From Theorem 1 and Propositions 1 and 2 we can easily derive the form of \mathcal{T}_ϕ^\perp , as follows.

Corollary 1 Under the assumptions of Theorem 1,

$$\mathcal{T}_\phi^\perp = \{[h(x) - E(h(X)|\sigma_\theta(X))_x][\tau(x, y, \theta) - E(\tau(X, Y, \theta)|X)_x] : h \in L_2(\lambda_0)\}.$$

We now compute the efficient score, which is the projection of the true score with respect to θ on to \mathcal{T}_ϕ^\perp . Let

$$r_0(x, y, \theta) = 2[\phi_0^{1/2}(x, y, \theta) - \phi_0^{1/2}(x, y, \theta_0)]/\phi_0^{1/2}(x, y, \theta_0).$$

The true score for the parameter of interest is the Fréchet derivative

$$\partial r_0(x, y, \theta)/\partial \theta|_{\theta=\theta_0}.$$

To differentiate from $\dot{r}_0(x, y, \theta_0)$, we denote the above derivative by $\overset{\circ}{r}_0(x, y, \theta_0)$. This is an $s(d-s)$ -dimensional vector. Since the mapping $T_x : L_2(\eta_0(x, \cdot, \theta)) \rightarrow \mathbb{R}$ and R_x also depend on θ , we now write them as $T_{x,\theta}$ and $R_{x,\theta}$. We use T_x to denote the mapping T_{x,θ_0} . Following Bickel et al (1993, Chapter 3), we use $\Pi(f|\mathcal{A})$ to represent the projection of a function f on to a subspace \mathcal{A} of $L_2(\phi_0)$.

Theorem 2 Suppose the following conditions hold.

1. For each $x \in \Omega_X$ and θ in a neighborhood of θ_0 , the mapping

$$T_{x,\theta} : L_2(\eta_0(x, \cdot, \theta)) \rightarrow \mathbb{R}$$

is continuously Fréchet differentiable in a neighborhood of $0 \in L_2(\eta_0(x, \cdot, \theta))$. Let $\tau(x, y, \theta)$ be the Riesz representation of $\dot{T}_{x,\theta}(0)$ and $\tau_c(x, y, \theta)$ be its centered version $\tau(x, y, \theta) - E_\theta[\tau(X, Y, \theta)|X]_x$.

2. The function $\theta \mapsto r_0(x, \cdot, \theta)$ is Fréchet differentiable at $\theta = \theta_0$ with Fréchet derivative $\overset{\circ}{r}_0(x, \cdot, \theta_0)$.

3. If

$$\begin{aligned} q_1(x, \theta_0) &= E[\overset{\circ}{r}_0(X, Y, \theta_0)\tau_c(X, Y, \theta_0)|X]_x \\ q_2(x, \theta_0) &= E[\tau_c^2(X, Y, \theta_0)|X]_x \\ q_3(x, \theta_0) &= q_1(x, \theta_0) - E_{\theta_0}[q_1(X, \theta_0)q_2^{-1}(X, \theta_0)|\sigma_{\theta_0}(X)]_x/E[q_2^{-1}(X, \theta_0)|\sigma_{\theta_0}(X)]_x \\ q_4(x, \theta_0) &= q_2^{-1}(x, \theta_0)q_3(x, \theta_0), \end{aligned}$$

where $q_2^{-1}(x, \theta_0)$ is the reciprocal of $q_2(x, \theta_0)$, then $q_4(x, \theta_0) \in L_2(\lambda_0)$.

Then

$$S_{\text{eff}}(x, y, \theta_0) = \Pi(\overset{\circ}{r}_0(x, y, \theta_0)|\mathcal{F}_\phi^\perp) = q_4(x, \theta_0)\tau_c(x, y, \theta_0). \quad (8)$$

PROOF. Let $u^*(x, y, \theta_0) = q_4(x, \theta_0)\tau_c(x, y, \theta_0)$. By the projection theorem, it suffices to show

- (a) $u^*(\cdot, \cdot, \theta_0) \in \mathcal{F}_\phi^\perp$;
- (b) for any $u \in \mathcal{F}_\phi^\perp$,

$$\langle \overset{\circ}{r}_0(\cdot, \cdot, \theta_0), u \rangle_{\phi_0} = \langle u^*(\cdot, \cdot, \theta_0), u \rangle_{\phi_0}. \quad (9)$$

By Condition 3, $q_4(\cdot, \theta_0) \in L_2(\lambda_0)$. Moreover, by the definition of q_4 in Condition 3 it is easy to verify that $E(q_4(X, \theta)|\sigma_{\theta_0}(X)) = 0$. Hence, assertion (a) holds.

Because $u \in \mathcal{F}_\phi^\perp$, it has the form $h(x, \theta_0)\tau_c(x, y, \theta_0)$ for some $h(\cdot, \theta_0) \in L_2(\lambda_0)$ satisfying $E[h(X, \theta_0)|\sigma_{\theta_0}(X)] = 0$. Hence, the right hand side of (9) is

$$\begin{aligned} E_\theta[h(X, \theta_0)q_4(X, \theta_0)\tau_c^2(X, Y, \theta_0)] &= E[h(X, \theta_0)q_4(X, \theta_0)q_2(x, \theta_0)] \\ &= E[h(X, \theta_0)q_3(X, \theta_0)]. \end{aligned}$$

Substitute the definition of $q_3(x, \theta_0)$ into the right hand side, and it becomes

$$E_\theta\{h(X, \theta)\{q_1(x, \theta_0) - E[q_1(x, \theta_0)q_2^{-1}(x, \theta_0)|\sigma_{\theta_0}(X)]/E[q_2^{-1}(x, \theta_0)|\sigma_{\theta_0}(X)]\}\}.$$

However, because $E(h(X, \theta_0)|\sigma_{\theta_0}(X)) = 0$, the equation above reduces to

$$E[h(X, \theta_0)q_1(X, \theta_0)] = E[h(X, \theta_0)\overset{\circ}{r}_0(X, Y, \theta_0)\tau_c(X, Y, \theta_0)],$$

which is the left hand side of (9). \square

The next corollary, which follows directly from Theorem 2, gives the general form for the efficient information estimating $\mathcal{S}_{T(Y|X)}$.

Corollary 2 *Under the assumptions of Theorem 2, the efficient information for estimating $\mathcal{S}_{T(Y|X)}$ is given by*

$$J_{\text{eff}}(\theta_0) = E[q_3(X, \theta_0)q_3^T(X, \theta_0)q_2^{-1}(X, \theta_0)]. \quad (10)$$

In the next three sections we apply the general result in Theorem 2 to derive the explicit forms of the efficient scores for three types of commonly used statistical functionals: the linear functionals, the composite linear functionals, and the implicit functionals. The common thread that runs through these developments is the calculation of the Riesz representation $\tau(x, y, \theta_0)$ of the Fréchet derivative $\dot{T}_x(0)$.

5 Linear statistical functionals

Dimension reduction of this type is the direct generalizations of the central mean subspace (Cook and Li, 2002) and the central moment subspace (Yin and Cook, 2002). It can also be viewed as a generalization of the single- and multiple-index models (see, for example, Härdle, Hall, and Ichimura, 1993). Let $f : \Omega_Y \rightarrow \mathbb{R}$ be a square-integrable function. Let L be the functional

$$L : \mathcal{G} \rightarrow \mathbb{R}, \quad g \mapsto \int_{\Omega_Y} f(y)g(y)d\mu_Y(y).$$

The corresponding conditional statistical functional is

$$L_{x,\theta}(r(x, \cdot, \theta)) \equiv L \circ R_{x,\theta}^{-1}(r(x, \cdot, \theta)) = \int_{\Omega_Y} f(y)(1 + r(x, y, \theta)/2)^2 \eta_0(x, y, \theta) dy.$$

The L -central subspace is defined by the relation

$$E[f(Y)|X] = E[f(Y)|\sigma_{\theta_0}(X)]. \quad (11)$$

Theorem 3 *Suppose the conditions 1, 2, 3 in Theorem 2 are satisfied for $L_{x,\theta}$. Then the efficient score for θ in problem (11) is given by (8) in which*

$$\begin{aligned} \tau_c(x, y, \theta_0) &= f(y) - E_{\theta_0}[f(Y)|\sigma_{\theta_0}(X)], \\ q_1(x, \theta_0) &= \partial E_{\theta}[f(Y)|\sigma_{\theta}(X)]_x / \partial \theta|_{\theta=\theta_0}, \\ q_2(x, \theta_0) &= E_{\theta_0}[f^2(Y)|X]_x - E_{\theta_0}^2[f(Y)|\sigma_{\theta_0}(X)]_x. \end{aligned}$$

PROOF. Because L_x is Fréchet differentiable at 0, its Fréchet derivative is the same as the Gâteaux derivative (Bickel et al, 1993, page 455), which is defined by

$$r(x, \cdot, \theta_0) \mapsto \partial L_x(\epsilon r(x, \cdot, \theta_0)) / \partial \epsilon|_{\epsilon=0}.$$

However, because

$$\begin{aligned}\partial L_x(\epsilon r(x, \cdot, \theta_0))/\partial \epsilon|_{\epsilon=0} &= \partial[\int_{\Omega_Y} f(y)(1 + \epsilon r(x, y, \theta_0)/2)^2 \eta_0(x, y, \theta_0) dy]/\partial \epsilon|_{\epsilon=0} \\ &= \int_{\Omega_Y} f(y) r(x, y, \theta_0) \eta_0(x, y, \theta_0) dy = \langle f, r(x, \cdot, \theta_0) \rangle_{\eta_0(x, \cdot, \theta_0)},\end{aligned}$$

the Riesz representation of $\dot{T}_x(0)$ is f . Hence, by Theorem 2,

$$q_2(x, \theta_0) = E[\tau_c^2(X, Y, \theta_0)|X]_x = E[f^2(Y)|X]_x - E^2[f(Y)|\sigma_{\theta_0}(X)]_x.$$

Also, for each θ ,

$$\begin{aligned}\int f(y)(1 + r_0(x, y, \theta)/2)^2 \eta_0(x, y, \theta) d\mu_Y(y) &= \int f(y) \eta_0(x, y, \theta) d\mu_Y(y) \\ &= E_\theta[f(Y)|X]_x = E_\theta[f(Y)|\sigma_\theta(X)]_x.\end{aligned}$$

Take Fréchet derivative with respect to θ on both sides to obtain

$$q_1(x, \theta_0) = E[f(Y)\dot{r}_0(X, Y, \theta_0)|X]_x = \partial E_\theta[f(Y)|\sigma_\theta(X)]_x/\partial \theta|_{\theta=\theta_0},$$

as desired. \square

Example 2 The central mean subspace introduced by Cook and Li (2002) is a special case of the L -central subspace with $f(y) = y$. The efficient score and efficient information are given by (8) and (10) where

$$\begin{aligned}\tau_c(x, y, \theta_0) &= y - E_{\theta_0}[Y|\sigma_{\theta_0}(X)]_x, \\ q_1(x, \theta_0) &= \partial E_\theta[Y|\sigma_\theta(X)]_x/\partial \theta|_{\theta=\theta_0}, \\ q_2(x, \theta_0) &= \text{var}_{\theta_0}(Y|X)_x = E_{\theta_0}(Y^2|X) - E_{\theta_0}^2[Y|\sigma_{\theta_0}(X)].\end{aligned}\tag{12}$$

For example, if the central mean subspace has dimension 1 and is spanned by $c + \Gamma\theta_0$ for some $c \in \mathbb{R}^p$ and $\Gamma \in \mathbb{R}^{p \times (p-1)}$, as described in the second paragraph of Section 2, then

$$\begin{aligned}\tau_c(x, y, \theta_0) &= y - E_{\theta_0}(Y|X^1 + \theta_0^\top \Gamma^\top X)_x, \\ q_1(x, \theta_0) &= \Gamma^\top x [\partial E_\theta(Y|X^1 + \theta^\top \Gamma^\top X)_x/\partial(\theta^\top \Gamma^\top x)]|_{\theta=\theta_0}.\end{aligned}$$

Therefore, the efficient score is

$$\begin{aligned}S_{\text{eff}}(x, y, \theta) &= \frac{1}{\text{var}_\theta(Y|X)_x} \frac{\partial E_\theta(Y|X^1 + \theta^\top \Gamma^\top X)_x}{\partial(\theta^\top \Gamma^\top x)} \\ &\quad \left\{ x - \frac{E_\theta[X/\text{var}_\theta(Y|X)|X^1 + \theta^\top \Gamma^\top X]_x}{E_\theta[1/\text{var}_\theta(Y|X)|X^1 + \theta^\top \Gamma^\top X]_x} \right\} [y - E_\theta(Y|X^1 + \theta^\top \Gamma^\top X)_x].\end{aligned}$$

The efficient information is

$$J_{\text{eff}}(\theta) = E_{\theta} \left[\frac{1}{\text{var}_{\theta}(Y|X)} \left(\frac{\partial E_{\theta}(Y|X^1 + \theta^{\top} \Gamma^{\top} X)}{\partial(\theta^{\top} \Gamma^{\top} X)} \right)^2 \left(X - \frac{E_{\theta}[X/\text{var}(Y|X)|X^1 + \theta^{\top} \Gamma^{\top} X]}{E_{\theta}[1/\text{var}_{\theta}(Y|X)|X^1 + \theta^{\top} \Gamma^{\top} X]} \right)^{\otimes 2} \right],$$

where $A^{\otimes 2}$ denotes AA^{\top} for a matrix A .

Alternatively, the efficient score and information can be written in the original parameterization β . See the External Appendix (Section II) for their explicit forms in the β -parameterization.

The central k th moment space (Yin and Cook, 2002) is a special case of the L -functional with $f(y) = y^k$. The efficient score and efficient information where (8) q_1 , q_2 , and τ_c given by formulas similar to (12) with Y replaced by Y^k .

Zhu and Zeng (2006) considered a dimension reduction problem defined through the characteristic function $E_{\theta_0}(e^{itY}|X) = E_{\theta_0}[e^{itY}|\sigma_{\theta_0}(X)]$. They used this relation to recover the central space, but if our goal is to estimate θ defined through this relation, then

$$\begin{aligned} \tau_c(x, y, \theta_0) &= e^{ity} - E_{\theta_0}[e^{itY}|\sigma_{\theta_0}(X)]_x, \\ q_2(x, \theta_0) &= \text{var}_{\theta_0}(e^{itY}|X)_x, \\ q_1(x, \theta_0) &= \partial E_{\theta}[e^{itY}|\sigma_{\theta}(X)]_x / \partial \theta|_{\theta=\theta_0}. \end{aligned}$$

The efficient score can be obtained by substituting the above into (8). \square

6 Composite linear statistical functionals

We now consider a nonlinear function of several linear functionals, which is motivated by dimension reduction for conditional variance considered in Zhu and Zhu (2009) and the single-index conditional heteroscedasticity model in Zhu, Dong, and Li (2012). See also Xia, Tong, and Li (2002). In fact, all cumulants are functionals of this type. Let T_1, \dots, T_k be bounded linear functionals from \mathcal{G} to \mathbb{R} . That is,

$$T_{\ell}(g) = \int f_{\ell}(y)g(y)d\mu_Y(y), \quad \ell = 1, \dots, k,$$

where f_1, \dots, f_k are square-integrable with respect to any density $g \in \mathcal{G}$. Let $\rho : \mathbb{R}^k \rightarrow \mathbb{R}$ be a differentiable function. Then $C : g \mapsto \rho(T_1(g), \dots, T_k(g))$ defines a statistical functional on \mathcal{G} to \mathbb{R} . We call such functionals *composite linear functionals*, or C-functionals. For example, if

$$T_1(g) = \int yg(y)d\mu_Y(y), \quad T_2(g) = \int y^2g(y)d\mu_Y(y), \quad \rho(s_1, s_2) = s_2 - s_1^2,$$

then $C(g) = \text{var}_g(Y)$ is the variance functional. The corresponding conditional statistical functional is defined by

$$C_{x,\theta}(r(x, \cdot, \theta)) = C \circ R_{x,\theta}^{-1}(r(x, \cdot, \theta_0)) = \rho[T_{1,x,\theta}(r(x, \cdot, \theta)), \dots, T_{k,x,\theta}(r(x, \cdot, \theta))]$$

where $T_{\ell,x,\theta}$ denotes $T_\ell \circ R_{x,\theta}^{-1}$. We will use the following notations:

$$\begin{aligned} \rho_\ell(X, \theta) &= \partial \rho(s) / \partial s_\ell |_{s=(T_{1,x,\theta}(0), \dots, T_{\ell,x,\theta}(0))}, \\ G(X, \theta) &= (\rho_1(X, \theta), \dots, \rho_k(X, \theta))^\top, \\ F(Y) &= (f_1(Y), \dots, f_k(Y))^\top. \end{aligned} \tag{13}$$

Also note that, in our case, $T_{\ell,x,\theta}(0) = E_\theta(f_\ell(Y)|X)_x$. Again, we use symbols such as $T_{\ell,x}$ and C_x to indicate T_{ℓ,x,θ_0} and C_{x,θ_0} .

Theorem 4 *Suppose the conditions 1, 2, 3 in Theorem 2 hold for $C_{x,\theta}$. Then the efficient score for $\mathcal{S}_{C(Y|X)}$ is given by (8), in which*

$$\begin{aligned} \tau_c(x, y, \theta_0) &= G^\top(x, \theta_0)F(y) - G^\top(x, \theta_0)E_{\theta_0}(F(Y)|X)_x, \\ q_1(x, \theta_0) &= \partial E_\theta(F^\top(Y)|X)_x / \partial \theta |_{\theta=\theta_0} G(x, \theta_0), \\ q_2(x, \theta_0) &= G^\top(x, \theta_0) \text{var}_{\theta_0}(F(Y)|X)_x G(x, \theta_0). \end{aligned} \tag{14}$$

PROOF. As shown in Section 5, the Riesz representation of $\dot{T}_{\ell,x}(0)$ is simply f_ℓ . By the chain rule of Fréchet differentiation and the definition (13), we have

$$\dot{C}_x(0) = \sum_{\ell=1}^k \rho_\ell(X, \theta_0) \dot{T}_{\ell,x}(0).$$

Hence, the Riesz representation of $\dot{C}_x(0)$ is

$$\tau(x, y, \theta_0) = \sum_{\ell=1}^k \rho_\ell(x, \theta_0) f_\ell(y) = G^\top(x, \theta_0)F(y).$$

In the meantime for each θ we have

$$\int \tau_c(x, y, \theta) \eta_0(x, y, \theta) d\mu_Y(y) = 0.$$

Differentiate both sides of this equation with respect to θ , to obtain

$$\int \partial \tau_c(x, y, \theta) / \partial \theta \eta_0(x, y, \theta) d\mu_Y(y) + \int \tau_c(x, y, \theta) \overset{\circ}{r}_0(x, y, \theta) d\mu_Y(y) = 0.$$

Hence,

$$\begin{aligned} \int \tau_c(x, y, \theta) \overset{\circ}{r}_0(x, y, \theta) d\mu_Y(y) &= -E_\theta[\partial \tau_c(X, Y, \theta) / \partial \theta | X] \\ &= -\partial G^\top(X, \theta) / \partial \theta [F(Y) - E_\theta(F(Y)|X)] + \partial E_\theta(F^\top(Y)|X) / \partial \theta G(X, \theta). \end{aligned}$$

Now take conditional expectation $E_\theta(\cdot \cdot | X)$ on both sides to prove the second relation in (14). \square

It is easy to see that an alternative expression of $q_1(\theta, X)$ in Theorem 4 is

$$q_1(x, \theta_0) = \partial \rho(E_\theta[f_1(Y)|X]_x, \dots, E_\theta[f_k(Y)|X]_x) / \partial \theta|_{\theta=\theta_0}.$$

This expression is useful because $\rho(E_\theta[f_1(Y)|X]_x, \dots, E_\theta[f_k(Y)|X]_x)$ is a function of $\sigma_\theta(X)$, and its derivative with respect to θ can be estimated by local linear regression, as we will see in Section 9.

Example 3 For the central variance subspace, we have

$$k = 2, f_1(y) = y, f_2(y) = y^2, \rho(s_1, s_2) = s_2 - s_1^2.$$

Hence, $F(y) = (y, y^2)^\top$, and

$$\begin{aligned} \rho_1(X, \theta_0) &= \partial(s_2 - s_1^2) / \partial s_1|_{s_1=E(F(Y)|X)} = -2E(Y|X), \\ \rho_2(X, \theta_0) &= \partial(s_2 - s_1^2) / \partial s_2|_{s_2=E(F(Y)|X)} = 1. \end{aligned}$$

The Riesz representation of $\dot{C}_x(0)$ and its centered version are

$$\begin{aligned} \tau(X, Y, \theta_0) &= -2E(Y|X)Y + Y^2, \\ \tau_c(X, Y, \theta_0) &= -2E(Y|X)Y + Y^2 - E[-2E(Y|X)Y + Y^2|X] \\ &= [Y - E(Y|X)]^2 - E[\text{var}(Y|X)|\sigma_{\theta_0}(X)]. \end{aligned}$$

Hence,

$$\begin{aligned} q_1(X, \theta_0) &= -\partial E_\theta[\tau_c(X, Y, \theta)] / \partial \theta|_{\theta=\theta_0} = \partial E_\theta[\text{var}_\theta(Y|X)|\sigma_\theta(X)] / \partial \theta|_{\theta=\theta_0}, \\ q_2(X, \theta_0) &= \text{var}[\tau_c(X, Y, \theta_0)|X] = \text{var}\{[Y - E(Y|X)]^2|X\}. \end{aligned}$$

In this case the subspace \mathcal{T}_ϕ^\perp consists of functions of the form

$$[h(x) - E(h(X)|\sigma_{\theta_0}(X))][Y - E(Y|X)]^2 - E(\text{var}(Y|X)|\sigma_{\theta_0}(X)),$$

where h is an arbitrary member of $L_2(\lambda_0)$. Interestingly, the estimating equation proposed by Zhu, Dong, and Li (2012) is a special member of \mathcal{T}_ϕ^\perp with $h(x)$ taken to be the components of x . \square

7 Implicit statistical functionals

In this section we study statistical functionals defined implicitly through estimating equations. Many robust estimators, such as conditional medians and quantiles, are of this type. Let $\Xi \subseteq \mathbb{R}$, and $e : \Xi \times \Omega_Y \rightarrow \mathbb{R}$ be a function of the parameter ξ and the variable y . Such functions are called estimating functions (see, for example, Godambe, 1960; Li and McCullagh, 1994). If the equation

$$\int_{\Omega_Y} e(\xi, y)g(y)d\mu_Y(y) = 0 \quad (15)$$

has a unique solution for each $g \in \mathcal{G}$, then it defines a functional $I : \mathcal{G} \rightarrow \mathbb{R}$ that assigns each g the solution to (15). We call such functionals *implicit functionals*, or I-functionals. If we replace $g \in \mathcal{G}$ by a conditional density function $\eta(\cdot, \cdot, \theta) \in \mathcal{H}$, then (15) becomes

$$\int_{\Omega_Y} e(\xi, y)\eta(x, y, \theta)d\mu_Y(y) = 0.$$

The corresponding conditional statistical functional is $I_{x,\theta}(r(x, \cdot, \theta)) = I \circ R_{x,\theta}^{-1}(r(x, \cdot, \theta))$. The I-central subspace is defined by the statement

$$I_X(0) \text{ is measurable with respect to } \sigma_{\theta_0}(X).$$

Naturally, we write the function $(\theta, x) \mapsto I_{x,\theta}(0)$ as $\xi(\theta, x)$.

To simplify the presentation we use the notion of generalized functions. Let \mathcal{K} be the class of functions defined on a bounded set B in \mathbb{R} that have derivatives of all orders, whose topology is defined by the uniform convergence of all derivatives. Any continuous linear functional $U : \mathcal{K} \rightarrow \mathbb{R}$ with respect to this topology is called a generalized function. For example, let $a \in B$. Then it can be shown that the linear functional

$$\delta_a : \mathcal{K} \rightarrow \mathbb{R}, \quad \phi \mapsto \phi(a)$$

is continuous with respect to this topology. This continuous linear functional is called the Dirac delta function. We identify the functional δ_a with an imagined function $x \mapsto \delta_a(x)$ on B and write $\delta_a(\phi)$ formally as the integral

$$\delta_a(\phi) = \int \phi(x)\delta_a(x)d\lambda(x).$$

A consequence of this convention is that if we pretend $\delta_a(x)$ to be the derivative $\partial I(x \leq a)/\partial a$ of the indicator function $I(x \leq a)$ then we get correct answers at the integral level. For example, for any constant a and small number ϵ , we have

$$\int [I(y \leq a + \epsilon) - I(y \leq a)]g(y)dy = \int \delta_a(y)\epsilon g(y)dy + o(\epsilon) = \epsilon g(a) + o(\epsilon).$$

Thus, the pretended linearization $I(y \leq a + \epsilon) - I(y \leq a) = \delta_a(y)\epsilon + o(\epsilon)$ has caused no inconsistency. We use this device to simplify our presentation of quantiles.

Theorem 5 Suppose the conditions 1, 2, 3 in Theorem 2 hold for $I_{x,\theta}$. Moreover, suppose for each $\xi \in \Xi$, $g \in \mathcal{G}$, there is a (generalized) function $\dot{e}(\xi, y)$, which plays the role of $\partial e(\xi, y)/\partial \xi$, such that

$$\left| \int_{\Omega_Y} [e(\xi + a, y) - e(\xi, y) - \dot{e}(\xi, y)a]g(y)d\mu_Y(y) \right| = o(a). \quad (16)$$

Then the efficient score for the I-central subspace is (8) in which

$$\begin{aligned} \tau_c(x, y, \theta_0) &= -e(\xi(\theta_0, x), y)/E_{\theta_0}[\dot{e}(\xi(\theta_0, x), Y)|X]_x, \\ q_1(x, \theta_0) &= \partial \xi(\theta, x)/\partial \theta|_{\theta=\theta_0}, \\ q_2(x, \theta_0) &= E_{\theta_0}[e^2(\xi(\theta, X), Y)|X]_x/E_{\theta_0}^2[\dot{e}(\xi(\theta_0, X), Y)|X]_x. \end{aligned} \quad (17)$$

PROOF. Differentiating both sides of the equation

$$\int_{\Omega_Y} e(I_{x,\theta}(er(x, \cdot, \theta)), y)(1 + er(x, y, \theta)/2)^2 \eta_0(x, y, \theta) d\mu_Y(y) = 0$$

with respect to ϵ at $\epsilon = 0$, and using the relation

$$\partial(1 + er(x, y, \theta)/2)^2/\partial \epsilon|_{\epsilon=0} = 2(1 + 0r(x, y, \theta)/2)r(x, y, \theta)/2 = r(x, y, \theta),$$

we find

$$\partial I_{x,\theta}(er(x, \cdot, \theta))/\partial \epsilon|_{\epsilon=0} = -E \left[\frac{e(I_{x,\theta}(0), Y)}{E(\dot{e}(I_{x,\theta}(0), Y)|X)_x} r(x, Y, \theta)|X \right]_x.$$

Hence, the Riesz representation of the Fréchet derivative $\dot{I}_{x,\theta}(0)$ is

$$\begin{aligned} \tau(x, y, \theta) &= -e(I_{x,\theta}(0), y)/E[\dot{e}(I_{x,\theta}(0), Y)|X]_x \\ &= -e(\xi(\theta, x), y)/E[\dot{e}(\xi(\theta, x), Y)|X]_x = \tau_c(x, y, \theta), \end{aligned} \quad (18)$$

where the last equality holds because, by definition, $E[e(\xi(\theta, X), Y)|X] = 0$. By (18) and the definition of q_1 in Theorem 2,

$$q_1(x, \theta_0) = -E[e(\xi(\theta_0, x), y)\dot{r}_0(x, y, \theta_0)|X]_x/E[\dot{e}(\xi(\theta_0, x), Y)|X]_x. \quad (19)$$

To further simplify the numerator of the right hand side, differentiate both sides of the equation $\int e(\xi(\theta, x), y)\eta_0(x, y, \theta)d\mu_Y(y) = 0$ to obtain

$$\begin{aligned} & \left[\int \dot{e}(\xi(\theta_0, x), y) \eta_0(x, y, \theta_0) d\mu_Y(y) \right] \dot{\xi}(\theta_0, x) \\ & + \int e(\xi(\theta_0, x), y) \dot{r}_0(x, y, \theta_0) \eta_0(x, y, \theta_0) d\mu_Y(y) = 0, \end{aligned}$$

where $\dot{\xi}(\theta_0, x)$ denotes $\partial \xi(\theta, x)/\partial \theta|_{\theta=\theta_0}$. Hence,

$$E[e(\xi(\theta_0, X), Y)\dot{r}_0(X, Y, \theta_0)|X]_x = -E[\dot{e}(\xi(\theta_0, X), Y)|X]_x \dot{\xi}(\theta_0, x).$$

Substitute this into (19) to prove the first relation in (17). Substitute (18) into the definition of q_2 in Theorem 2 to prove the second relation in (17). \square

In the next example we derive the efficient score for a particular type of I-functional — the quantile functional.

Example 4 If we assume all densities in \mathcal{G} are continuous, then the p th quantile is the solution to the equation $EI(Y \leq \xi) = p$. Equivalently, ξ is the solution to the equation

$$E[e(\xi, y)] = E[-\text{sgn}(y - \xi) + 1 - 2p] = 0.$$

Because the generalized derivative of $\text{sgn}(t)$ is $2\delta_0(t)$, we have $\dot{e}(\xi, y) = 2\delta_\xi(y)$. Hence

$$E[\dot{e}(\xi(\theta, X), Y)|X]_x = \int_{\Omega_Y} 2\delta_{\xi(\theta, x)}(y)\eta_0(x, y, \theta)dy = 2\eta_0(x, \xi(\theta, x), \theta).$$

Because $e(\xi(\theta, x), Y)$ is a binary random variable that takes the value $2(1-p)$ with probability p and $-2p$ with probability $(1-p)$, we have

$$E[e^2(\xi(\theta, x), Y)|X]_x = [2(1-p)]^2p + (-2p)^2(1-p) = 4(1-p)p.$$

Hence, in the efficient score,

$$q_2(x, \theta_0) = \eta_0^2(x, \xi(\theta_0, x), \theta_0)/[(1-p)p],$$

and $q_1(x, \theta_0)$ is as given in (17) with $\xi(\theta, x)$ being the conditional p th quantile of Y given X . \square

8 Effect of estimating the central subspace

Throughout the previous sections we have treated X as the true predictor from the central subspace; that is, $X = \zeta^\top \tilde{X}$ where \tilde{X} is the original predictor and ζ is a basis matrix of the central subspace based for $Y|\tilde{X}$. However, in practice, ζ itself needs to be estimated and, in theory at least, should affect the form of the efficient score about β . While our simulation studies in Section 10 indicate that this effect is very small, for theoretical rigor we present here the efficient score treating the central subspace as an additional (finite-dimensional) nuisance parameter. For convenience, we use ζ to denote both a $p \times d$ matrix and the corresponding $(p-d)d$ -dimensional Grassmann manifold.

Let $S_{\text{eff}}(\zeta^\top \tilde{X}, Y, \theta)$ denote the efficient score in Theorem 2 with X replaced by $\zeta^\top \tilde{X}$. Let $S_{\text{eff}}^*(\tilde{X}, Y, \zeta, \theta)$ denote the efficient score for θ with ζ treated as an

additional nuisance parameter. Let $s_\zeta(\tilde{x}, y, \zeta, \theta)$ denote the score with respect to ζ . In the External Appendix (Section III) it is shown that

$$S_{\text{eff}}^* = S_{\text{eff}} - \Pi(S_{\text{eff}} | \text{span}(\Pi(s_\zeta | \mathcal{F}_\phi^\perp))) \equiv S_{\text{eff}} - g, \quad (20)$$

where g is the function

$$g = q_2^{-1}(\zeta_0^\top \tilde{x}, \theta_0) \tau_c(\zeta_0^\top \tilde{x}, y, \theta_0) E\{q_3(\zeta_0^\top \tilde{X}, \theta_0) q_3^{*\top}(\zeta_0^\top \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\top \tilde{x}, \theta_0)\} \\ E\{q_3^*(\zeta_0^\top \tilde{X}, \theta_0) q_3^{*\top}(\zeta_0^\top \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\top \tilde{x}, \theta_0)\}^\dagger q_3^*(\zeta_0^\top \tilde{x}, \theta_0),$$

in which \dagger indicates the Moore-Penrose inverse, and

$$q_1^*(\zeta_0^\top \tilde{x}, \theta_0) = E[s_\zeta(\tilde{X}, Y, \zeta_0, \theta_0) \tau_c(\zeta_0^\top \tilde{X}, Y, \theta_0) | \zeta_0^\top \tilde{X}]_{\tilde{x}} \\ q_3^*(\zeta_0^\top \tilde{x}, \theta_0) = q_1^*(\zeta_0^\top \tilde{x}, \theta_0) \\ - E_{\theta_0}[q_1^*(\zeta_0^\top \tilde{X}, \theta_0) q_2^{-1}(\zeta_0^\top \tilde{x}, \theta_0) | \sigma_{\theta_0}(\zeta_0^\top \tilde{X})]_{\tilde{x}} / E[q_2^{-1}(\zeta_0^\top \tilde{x}, \theta_0) | \sigma_{\theta_0}(\zeta_0^\top \tilde{X})]_{\tilde{x}}.$$

In theory, the asymptotic variance bound based on S_{eff} is lower than or equal to that based on S_{eff}^* . However, in the simulation studies (Table 2 in Section 10) we see that the theoretical lower bound based on S_{eff} is nearly reached, which indicates that effect of estimating the central subspace on the efficient score for β is very small.

9 Estimation

In this section we introduce semiparametrically efficient estimators using the theory developed in the previous sections. For the L-functionals we develop the estimator in full generality, but for the C- and I-functionals we focus on the conditional variance functional and the conditional quantile functional. Procedures for other C- or I-functionals can be developed by analogy.

We first clarify two points related to the algorithm we will propose. First, since we will rely heavily on the MAVE-type algorithms, it is more convenient to use the β -parameterization rather than the θ -parameterization, and avoid redundancy in β by taking the generalized matrix inverse. We will justify the β -parameterization after introducing the algorithm. Second, the MAVE algorithm actually has two variants: the outer product gradient (OPG) and a refined version of MAVE (RMAVE). Typically, OPG, MAVE, and RMAVE are progressively more accurate and require more computation. In the following, the MAVE-type algorithm can be replaced either by RMAVE for greater accuracy or by OPG for less computation.

Our estimation procedure is divided into four steps. In step 1, we estimate the central subspace and project \tilde{X} on to this subspace to obtain X . In step 2, we estimate $T(\eta_0(X_1, \cdot)), \dots, T(\eta_0(X_n, \cdot))$ using a d -dimensional kernel estimate. These estimates are used as the proxy response, and we denote them by $\hat{Y}_1, \dots, \hat{Y}_n$.

In step 3, we apply MAVE to $(X_1, \hat{Y}_1), \dots, (X_n, \hat{Y}_n)$ to estimate an initial value for β . In step 4, we use one-step Newton-Raphson algorithm based on the efficient score and efficient information to approximate the semiparametrically efficient estimate. We call our estimator SEE, which stands for *semiparametrically efficient estimator*.

Preparation step: a MAVE code. Let $(X_1, U_1), \dots, (X_n, U_n)$ be a random sample from (X, U) , and $K_h(\cdot)$ be a kernel with bandwidth h . That is, $K_h(t) = K(t/h)/h$ for some symmetric function K that integrates to 1 and $h > 0$. Compute the objective function

$$\Gamma(a, b, A) = \sum_{i=1}^n \sum_{j=1}^n [(U_j - a_i - b_i^\top A^\top (X_j - X_i))^2 K_h(X_j - X_i)] \quad (21)$$

where $a_1, \dots, a_n \in \mathbb{R}$, $b_1, \dots, b_n \in \mathbb{R}^d$, a and b on the left denote $(a_1, \dots, a_n)^\top$ and $(b_1^\top, \dots, b_n^\top)^\top$ respectively, and $A \in \mathbb{R}^{p \times d}$. We will use this objective function in several ways. It is well known that minimization of (21) can be solved by iterations of least squares, and in each iteration there is an explicit solution. Thus purely search-based numerical optimization (such as the simplex method) is avoided. See, for example, Li, Li, and Zhu (2010) and Yin and Li (2011).

Step 1: estimation of central subspace. We use the MAVE-ensemble in Yin and Li (2011) to estimate the central subspace. In this procedure, U in (21) is taken to be a set of functions $\{f_1(Y), \dots, f_m(Y)\}$ randomly sampled from a dense family in $L_2(\mu_Y)$. In this paper take this set to be $\{(\sin(t_i y), \cos(t_i y)) : i = 1, \dots, 10\}$, where t_1, \dots, t_{10} are i.i.d. $\text{unif}(0, 4)$. The sample of responses Y_1, \dots, Y_n are standardized so that the range $(0, 4)$ of the uniform distribution represents a reasonably rich class of functions relative to the range of Y . The basis matrix ζ of $\mathcal{S}_{Y|X}$ is then estimated by the MAVE-ensemble. The projected predictor $X = \hat{\zeta}^\top \tilde{X}$ is taken as the predictor in steps 2–4. Since our goal is to estimate $\mathcal{S}_{T(Y|X)}$, the choice of the working dimension \hat{d} of $\mathcal{S}_{Y|X}$ is not crucial.

As was shown in Yin and Li (2011), at the population level, the MAVE ensemble is guaranteed to recover central subspace exhaustively as long as the functions of Y form a characterizing family. In practice, it is true any information lost in the initial step will be inherited by SEE. However, our experiences indicate that this problem can be mitigated by using a sufficiently rich ensemble family — for example, by increasing the range of the uniform distribution and the number of t_i 's.

Several other methods are available for exhaustive estimation of the central subspace, such as the the semiparametrically efficient estimator of Ma and Zhu (2013a), the DMAVE by Xia (2007), and the Sliced Regression by Wang and Xia (2008). Here, we have chosen the MAVE-ensemble for its computational simplicity.

Step 2: estimation of proxy response. This step is unnecessary for the L-functionals: because $T(\eta(X, \cdot)) = E[f(Y)|X]$ for some function f , we can use $f(Y)$ itself as the proxy response \hat{Y} . For the conditional variance functionals, this step needs not be fully performed: we can use $(Y - \hat{E}(Y|X))^2$ as the proxy response \hat{Y} , where $\hat{E}(Y|X)$ is the kernel estimator of $E(Y|X)$. If simplification of this type is not applicable, then we need to perform nonparametric estimation of $T(\eta(X, \cdot))$. For example, for the I-functionals, we use the minimizer ξ_i^* of the function $E_n[e(Y, \xi)K_h(X - X_i)]$ as the proxy response \hat{Y}_i .

Step 3: initial estimate of β . Apply MAVE to $(X_1, \hat{Y}_1), \dots, (X_n, \hat{Y}_n)$ to obtain an initial estimate of β by minimizing $\Gamma(a, b, \beta)$ over all a, b, β . Denote this initial estimate as $\tilde{\beta}$.

Step 4: the one-step Newton-Raphson algorithm. Rather than attempting to solve the score equation (4), we propose a one-step Newton-Raphson procedure. Let $S_{\text{eff}}(X, Y, \beta)$ be the efficient score in β , which is obtained by replacing $\sigma_\theta(X)$ by $\beta^\top X$ and $\partial/\partial\theta$ by $\partial/\partial\text{vec}(\beta)$ whenever applicable. Let $J_{n,\text{eff}}(\beta) = E_n[S_{\text{eff}}(X, Y, \beta)S_{\text{eff}}^\top(X, Y, \beta)]$. We estimate β by

$$\hat{\beta} = \tilde{\beta} + [J_{n,\text{eff}}(\tilde{\beta})]^\dagger E_n[S_{\text{eff}}(X, Y, \tilde{\beta})] \quad (22)$$

where $\tilde{\beta}$ is the initial value from Step 3.

We now describe in detail how to compute $E_n[S_{\text{eff}}(X, Y, \tilde{\beta})]$ for different functionals. For the L-functional, the efficient score involves the following functions

$$(a) \quad E[f(Y)|\beta^\top X], \quad (b) \quad E[f^2(Y)|X], \quad (c) \quad \partial E[f(Y)|\beta^\top X]/\partial\text{vec}(\beta).$$

For the C-functional, it involves the functions:

$$(d) \quad E_{\theta_0}[f_\ell(Y)|X], \quad (e) \quad \partial\rho(E_\beta[f_1(Y)|X], \dots, E_\beta[f_\ell(Y)|X])/\partial\text{vec}(\beta), \\ (f) \quad \text{cov}[f_i(Y), f_j(Y)|X].$$

For the conditional quantile functional, it involves the functions

$$(g) \quad E[\xi(X)|\beta^\top X], \quad (h) \quad \partial E[\xi(X)|\beta^\top X]/\partial\text{vec}(\beta), \quad (i) \quad \eta_0(X, E[\xi(X)|\beta^\top X]).$$

These functions can be categorized into three types: (a), (b), (d), (f), (g) are conditional means of random variables; (c), (e), (h) are derivatives of functions of $\beta^\top X$ with respect to $\text{vec}(\beta)$; (i) is the conditional density evaluated at a quantile. The first two types can be solved by minimizing $\Gamma(a, b, A)$ in (21) with specific U_i and A : the following table gives the details of these random variables and matrices, as well as which parts of the MAVE output are needed in SEE.

Table 1: Using MAVE to estimate quantities (a) through (h) in efficient score

quantities	A	U_i	MAVE output
(a)	β	$f(Y)$	a^*
(b)	I_p	$f^2(Y)$	a^*
(c)	β	$f(Y)$	b^*
(d)	I_p	$f_\ell(Y)$	a^*
(e)	β	$\rho(E_n[f_1(Y) X], \dots, E_n[f_k(Y) X])$	b^*
(f)	I_p	$f_\ell(Y), f_\ell(Y)f_{\ell'}(Y)$	a^*
(g)	β	$\xi_n(X)$	a^*
(h)	β	$\xi_n(X)$	b^*

Finally, we estimate $\eta_0(x, y_0)$ for any y_0 by the kernel conditional density estimator:

$$E_n[K_{h_1}(Y - y_0)K_h(X - x)]/E_n[K_h(X - x)],$$

where h_1 is a different bandwidth for the response variable.

Tuning of bandwidths. We need to determine the kernel bandwidths h at various stages in the above algorithm. We use the Gaussian kernel with optimal bandwidth $h = cn^{-1/(p+4)}$ for nonparametric regression (see, for example, Xia, Tong, Li, and Zhu, 2002), where c is determined by five-fold cross validation. That is, we randomly divide the data into 5 subsets of roughly equal sizes. For each $i = 1, \dots, 5$, we use the i th subset as the testing set and the rest as the training set. For a given c , we conduct dimension reduction on the training set, and use the sufficient predictor to evaluate a certain prediction criterion at each point on the testing set and average these evaluations over the testing set, and finally average the five averages of the criterion to obtain a single number. We then minimize the resulting criterion over c by a grid search. Naturally, the prediction criterion depends on the object to be estimated using that kernel. Below is a list of the prediction criteria we propose for the four steps in the estimation procedure.

In Step 1: We use the distance correlation introduced by Székely and Rizzo (2009, Theorem 1, expression (2.11); p and q therein are taken to be 2).

In Step 2: For the L-functionals, no tuning is needed. For the conditional variance functional, we need to estimate $E(Y|X)$, and we use the prediction criterion $[Y - \hat{E}(Y|X)]^2$, where $\hat{E}(Y|X)$ is the kernel estimate of $E(Y|X)$ based on the training set using a tuning constant c . For conditional median, we use the prediction criterion $|Y - \hat{M}(Y|X)|$, where $\hat{M}(Y|X)$ is the kernel estimate of the conditional median based on the training set using a specific tuning constant.

In Step 3: We use the prediction criterion $[\hat{Y} - \hat{E}(\hat{Y}|\tilde{\beta}^\top X)]^2$, where \hat{Y} is the proxy response obtained from Step 2, and $\hat{E}(\hat{Y}|\tilde{\beta}^\top X)$ is the MAVE output a^* .

In Step 4: There are three types of kernels in this step: the kernel for X , the kernel for $\tilde{\beta}^\top X$, and the kernel for Y (the last one is needed only for the conditional median functional). Corresponding to these types we use bandwidths

$$h = cn^{-1/(d+4)}, \quad h = cn^{-1/(s+4)}, \quad h = cn^{-1/(1+4)}.$$

We use cross validation to determine the common c . Once again, we use different prediction criteria for different functionals. For the conditional mean functional, we use the criterion $[Y - \hat{E}(Y|\hat{\beta}^\top X)]^2$. For conditional variance functional, we use the criterion $\{(Y - \hat{E}(Y|X))^2 - \hat{E}[Y - \hat{E}(Y|X)|\hat{\beta}^\top X]^2\}^2$. For the conditional median functional, we use the criterion $|Y - \hat{E}[\hat{M}(Y|X)|\hat{\beta}^\top X]|$.

Justification of parameterization. We now justify the one-step iteration formula (22) as an equivalent form of

$$\hat{\theta} = \tilde{\theta} + J_{n,\text{eff}}(\tilde{\theta})^{-1} E_n[S_{\text{eff}}(X, Y, \tilde{\theta})]. \quad (23)$$

Since β is a function of a $s(d-s)$ dimensional parameter θ , the efficient information $J_{n,\text{eff}}(\beta)$ has rank $s(d-s)$. Let Γ denote the $sd \times [s(d-s)]$ matrix whose columns are eigenvectors of $J_{n,\text{eff}}(\beta)$ corresponding to its nonzero eigenvalues, and let $\beta = \Gamma\theta$, where θ is a free parameter in $\mathbb{R}^{s(d-s)}$. In this parameterization,

$$S_{\text{eff}}(X, Y, \theta) = \Gamma^\top S_{\text{eff}}(X, Y, \beta), \quad J_{n,\text{eff}}(\theta) = \Gamma^\top J_{n,\text{eff}}(\beta)\Gamma.$$

Hence (23) is equivalent to

$$\hat{\theta} = \tilde{\theta} + [\Gamma^\top J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1} \Gamma^\top E_n[S_{\text{eff}}(X, Y, \tilde{\beta})].$$

Multiply both sides by Γ from the left, we find

$$\Gamma\hat{\theta} = \Gamma\tilde{\theta} + \Gamma[\Gamma^\top J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1} \Gamma^\top E_n[S_{\text{eff}}(X, Y, \tilde{\beta})].$$

By construction, $\Gamma\hat{\theta} = \hat{\beta}$, $\Gamma\tilde{\theta} = \tilde{\beta}$, and $\Gamma[\Gamma^\top J_{n,\text{eff}}(\tilde{\beta})\Gamma]^{-1} \Gamma^\top$ is Moore-Penrose inverse of $J_{n,\text{eff}}(\tilde{\beta})$. Thus the above iterative formula is the same as (22).

In concluding this section we point out two attractive features of our algorithm, which were briefly touched on in the Introduction. First, since our algorithm is implemented by repeated applications of variations of MAVE, it essentially consists of sequence of least squares algorithms, thus avoiding any search-based numerical optimization, which can be infeasible when the dimension of θ is high. The second advantage is that, since our algorithm is based on the β -parameterization, we do not need any subjectively chosen parameterization. In comparison, Ma and Zhu (2013a) used the parameterization $\beta = (I_s, \theta)^\top$, where $\theta \in \mathbb{R}^{(d-s) \times s}$ is a matrix with free-varying entries. Note that this is not without loss of generality, because in reality the first d rows of β can be linearly dependent.

10 Simulation comparisons

In this section we conduct simulation comparisons between SEE and other methods for estimating three types of T -central subspaces: conditional mean, conditional variance, and conditional quantile. We use the distance between two subspaces proposed by Li, Zha, and Chiaromonte (2005) to measure estimation errors, which is defined as

$$\text{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|\Pi_{\mathcal{S}_1} - \Pi_{\mathcal{S}_2}\|_2, \quad (24)$$

where \mathcal{S}_1 and \mathcal{S}_2 are subspaces of \mathbb{R}^p , and $\|\cdot\|_2$ is the L_2 norm in $\mathbb{R}^{p \times p}$. For each of the following models, the sample size is taken to be $n = 200$ or 500 , or both; each sample is repeatedly drawn for $n_{\text{sim}} = 100$ times in the simulation. In all simulations we fix the working dimension in Step 1 at $\hat{d} = 3$, even though d is no greater than 2 in all examples.

The explicit forms of the efficient scores efficient information for Models I \sim VII used in the following comparisons are derived in the External Appendix (Section II).

(a) Comparison for the central mean subspace. In this case the functional $T(\eta(X, \cdot))$ is the conditional mean $E(Y|X)$. We compare SEE with RMAVE under the following models

$$\begin{aligned} \text{Model I: } Y &= X_1 + (1 + |X_2|)\varepsilon, \\ \text{Model II: } Y &= X_1(X_1 + X_2 + 1) + 0.5\varepsilon, \\ \text{Model III: } Y &= X_1 + (1 + |X_1|)\varepsilon, \\ \text{Model IV: } Y|X &\sim \text{Poisson}(|X_1 + X_2|), \end{aligned}$$

where $X \sim N(0, I_{10})$, $X \perp \varepsilon$, and $\varepsilon \sim N(0, 1)$. These models represent a variety of scenarios one might encounter in practice. Specifically, the central mean subspace is a proper subspace of the central subspace in Model I, but coincides with the latter in the other models. The conditional variance $\text{var}(Y|X)$ is a constant in Model II, but depends on X in the other models. Because of its additive error structure Model II is favorable to MAVE. Finally, model IV has a discrete response and the error only enters implicitly. Model I and Model III will be used again for Comparison 2, where conditional variance is the target; Model II was also used in Li (1991) and Xia, Tong, Li, and Zhu (2002).

The results with sample sizes $n = 200$ and $n = 500$ are presented in Table 2, in the blocks indicated by $E(Y|X)$. The entries are in the form $a(b)$, where a is the mean, and b the standard error, of the distance (24) between the true and estimated $\mathcal{S}_{E(Y|X)}$, based on $n_{\text{sim}} = 100$ simulated samples.

(b) Comparison for the central variance subspace. Let $T(\eta(X, \cdot))$ be the conditional variance $\text{var}(Y|X)$. We compare SEE with the estimator proposed in Zhu and Zhu (2009) and Zhu, Dong, and Li (2012). In Model I, the central variance subspace is different from either the central mean subspace or the central subspace, while in Model III, the three spaces coincide. The results with $n = 200$ and $n = 500$ are reported in Table 2, in the blocks indicated by $\text{var}(Y|X)$.

(c) Comparison for the central median subspace. Let $T(\eta(X, \cdot))$ be the conditional median $M(Y|X)$. We compare SEE with the adaptive quantile estimator (AQE) introduced by Kong and Xia (2012), which can also be used to estimate the central median subspace. We use the following models:

$$\text{Model V: } Y = X_1^2 + X_2\varepsilon, \quad \text{Model VI: } Y = 3X_1 + X_2 + \varepsilon,$$

where $X \sim N(0, I_{10})$ and $X \perp \varepsilon$. For model V, ε has a skewed-Laplace distribution with p.d.f.

$$f(\varepsilon) = \begin{cases} (5/4) e^{-5\varepsilon/2} & \varepsilon \geq -(2/5) \log(4/3) \\ (80/27) e^{5\varepsilon} & \varepsilon < -(2/5) \log(4/3) \end{cases}.$$

In this case,

$$E(Y|X) = X_1^2 + X_2[1/5 - 2/5 \log(4/3)], \quad M(Y|X) = X_1^2.$$

It follows that

$$\mathcal{S}_{E(Y|X)} = \text{span}\{(1, 0, \dots, 0)^\top, (0, 1, 0, \dots, 0)^\top\}, \quad \mathcal{S}_{M(Y|X)} = \text{span}\{(1, 0, \dots, 0)^\top\}.$$

For model VI, $\varepsilon \sim t_{(3)}$. Although for this model the central mean subspace coincides with the central median subspace, due to the heavy-tailed error distribution the conditional median is preferred to the conditional mean. Similar models can be found, for example, in Zou and Yuan (2008). The results for sample sizes $n = 200, 500$ are presented in Table 2, in the blocks indicated by $M(Y|X)$.

(d) Comparison with theoretical lower bound To see how closely the theoretical asymptotic lower bound (ALB) is approached by SEE for finite samples, we now compute the limit

$$\lim_{n \rightarrow \infty} \sqrt{n} E(\|\Pi_{\text{span}(\hat{\beta})} - \Pi_{\text{span}(\beta_0)}\|_2), \quad (25)$$

where $\hat{\beta}$ is the semiparametrically efficient estimate. This is the best we can do to estimate the T -central subspace. The explicit form and the derivation of (25) is given in the External Appendix (Section IV). We present the numerical values of this limit in the last column (under the heading ALB) of Table 2 for different models and T functionals.

(e) Conclusions for comparisons in (a) ~ (d) From Table 2 we see that SEE achieves substantially improved accuracy across all models and functionals considered. Stability of the estimates is also improved as can be seen from the decrease in standard errors. Our simulation studies (not presented here) indicate that the results are not significantly affected by the working dimension of the central subspace. For example, we repeated the analysis with $d = 2, 4$ and the patterns of the comparisons are not significantly altered.

Comparing the results for $n = 200$ and $n = 500$, we see that the proportion of improvement is smaller for the large sample size, as to be expected.

We see that the actual errors of the SEE computed from simulations are very close to the theoretical lower bounds both sample sized $n = 200, 500$ and the differences become negligible for $n = 500$. Since in the estimator the central subspace is estimated from the sample and in the lower bounds the central subspace is treated as known, the closeness of these errors to their corresponding ALB also indicates that the lower bound based on S_{eff}^* in Section 8 is close to the lower bound based on S_{eff} . In other words the effect of estimating the central subspace on the efficient score is small.

Table 2: Comparison of SEE with other estimators for three statistical functionals

n	Functionals	Models	Estimators			
200	$E(Y X)$		RMAVE		SEE	ALB
		I	0.519 (0.127)		0.153 (0.067)	0.175
		II	0.164 (0.063)		0.124 (0.054)	0.115
		III	0.490 (0.156)		0.165 (0.058)	0.147
		IV	0.206 (0.072)		0.078 (0.036)	0.071
	$\text{var}(Y X)$		Zhu-Zhu	Zhu-Dong-Li	SEE	ALB
		I	0.656 (0.231)	0.283 (0.126)	0.125 (0.051)	0.116
		III	0.843 (0.197)	0.408 (0.193)	0.116 (0.055)	0.113
	$M(Y X)$		AQE		SEE	ALB
		V	0.029 (0.012)		0.019 (0.013)	0.016
		VI	0.087 (0.022)		0.049 (0.019)	0.043
	500	$E(Y X)$		RMAVE		SEE
I			0.100 (0.030)		0.081 (0.021)	0.083
II			0.081 (0.018)		0.073 (0.013)	0.073
III			0.095 (0.033)		0.047 (0.015)	0.046
IV			0.079 (0.015)		0.047 (0.010)	0.045
$\text{var}(Y X)$			Zhu-Zhu	Zhu-Dong-Li	SEE	ALB
		I	0.315 (0.066)	0.219 (0.034)	0.109 (0.020)	0.104
		III	0.236 (0.035)	0.183 (0.037)	0.071 (0.025)	0.066
$M(Y X)$			AQE		SEE	ALB
		V	0.017 (0.005)		0.009 (0.003)	0.010
		VI	0.042 (0.015)		0.031 (0.009)	0.027

(f) Comparison under dependent components of X We now repeat comparisons in (a) through (d) using an X with dependent components. Rather than taking $\text{var}(X) = I_{10}$ we now take

$$\text{cov}(X^i, X^j) = 0.5^{|i-j|}, \quad i, j = 1, \dots, 10. \quad (26)$$

The same covariance matrix was used in Ma and Zhu (2012). The results parallel to those in Table 2 are presented in Table 3. We see that the errors are larger than those for X with independent components, but the degree by which SEE improves upon the other estimators, and to which it approaches theoretical asymptotic lower bound, are similar to those for the independent-component case.

Table 3: Comparison of SEE with other estimators with correlated predictors

Functionals	Models	Estimators			
$E(Y X)$		RMAVE		SEE	ALB
	I	0.520 (0.155)		0.164 (0.066)	0.168
	II	0.404 (0.165)		0.160 (0.080)	0.149
	III	0.571 (0.134)		0.304 (0.075)	0.289
	IV	0.283 (0.090)		0.075 (0.023)	0.085
$\text{var}(Y X)$		Zhu-Zhu	Zhu-Dong-Li	SEE	ALB
	I	0.539 (0.174)	0.431 (0.230)	0.131 (0.077)	0.108
	III	0.617 (0.222)	0.303 (0.169)	0.204 (0.066)	0.227
$M(Y X)$		AQE		SEE	ALB
	V	0.074 (0.023)		0.015 (0.012)	0.012
	VI	0.076 (0.061)		0.032 (0.015)	0.041

(g) Comparison for conditional upper quartile We now apply SEE to estimating the central upper-quartile subspace in which the functional of interest is solution to the equation $P(Y \leq c|X) = 0.75$. We generate X from $N(0, \Sigma)$ with Σ given by (26). We compare SEE with AQE for models V and VI, and the additional model

$$\text{Model VII: } Y = 1 + X_1 + (1 + 0.4 X_2) \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. In Model V, the central upper-quartile subspace has dimension 2, spanned by $(1, 0, \dots, 0)^\top$ and $(0, 1, 0, \dots, 0)^\top$; in Models VI and VII, the central upper-quartile subspaces have dimension 1 and are spanned by $(3, 1, 0, \dots, 0)^\top$ and $(1, 0.4\Phi^{-1}(0.25), 0, \dots, 0)^\top$, respectively, where Φ is the c.d.f. of the standard normal distribution. The performance of the estimators is summarized in Table 4.

We see that SEE outperforms AQE both in average accuracy and estimation stability. It is also interesting to note that, for Model VI, the central median subspace coincides with the central upper-quartile subspace, and the SEE based on the conditional median (Table 3) performs better than the SEE based on the conditional upper quartile (Table 4).

Table 4: Comparison of SEE with AQE at $\tau = 0.75$

Models	AQE	SEE	ALB
V	0.176 (0.083)	0.043 (0.017)	0.042
VI	0.160 (0.037)	0.069 (0.021)	0.053
VII	0.245 (0.117)	0.109 (0.072)	0.111

11 Application: age of abalones

In this section we evaluate the performance of SEE in an application, which is concerned with predicting the age of abalones using their physical measurements. The data can be found at the website <http://archive.ics.uci.edu/ml/datasets.html>, and consist of observations from 4177 abalones, of which 1528 are male, 1307 are female, and 1342 are infant. The observations on each subject contain 7 physical measurements and the age of the subject, as measured by the number of rings in its shell. We only use the subset of male abalones. The 532th subject in this subset is an outlier, and is deleted. Thus we have a sample of size 1527 with 7 predictors and 1 response. For objective evaluation of the estimators we further split the data into two subsets: the first 764 subjects are used as the training set to estimate the sufficient predictors and the rest 763 subjects are used as the testing set to plot the derived sufficient predictors versus the response.

We estimate both the central mean subspace (CMS) and the central variance subspace (CVS) of this data set. The CMS is estimated by RMAVE, SEE, and the method implicitly contained in Zhu, Dong, and Li (2012). The CVS is estimated by the methods proposed by Zhu and Zhu (2009) and Zhu, Dong, and Li (2012), and the SEE. The results are presented in Figure 1. The three upper panels are scatter plots of Y versus the sufficient predictor in the CMS as estimated by RMAVE, Zhu-Dong-Li, and SEE in that order. The three lower panels are the scatter plots of the absolute residuals $|Y - \hat{E}(Y|X)|$ versus the sufficient predictor in the CVS as estimated by Zhu-Zhu, Zhu-Dong-Li, and SEE.

To give an objective numerical comparison, we use a bootstrapped error measurement akin to that introduced by Ye and Weiss (2003), which is reasonable because all estimators involved are consistent. Since the predictors in the abalone data set are highly correlated, two estimates of β that span substantially different linear spaces can correspond to nearly identical $\beta^\top X$. For this reason, rather than measuring the error in $\hat{\beta}$, as we did in the simulations, here we directly measure the error in $\hat{\beta}^\top X$. Specifically, we generate 500 bootstrap samples, and for each sample we compute the estimate $\tilde{\beta}$. We also compute the full-sample estimate $\hat{\beta}$. For each bootstrap sample, we evaluate the sample correlation between

$$\{\tilde{\beta}^\top X_1, \dots, \tilde{\beta}^\top X_n\}, \quad \{\hat{\beta}^\top X_1, \dots, \hat{\beta}^\top X_n\}.$$

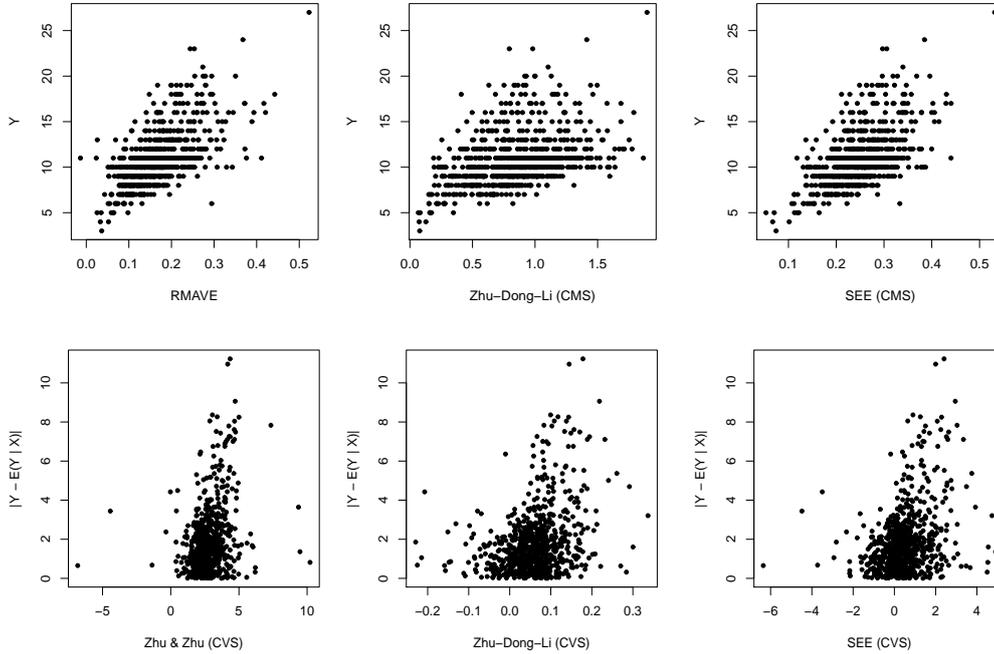


Figure 1: Comparison of SEE with other estimators of CMS and CVS for the abalone data. In the upper panels, the x-axes are the predictors obtained by RMAVE, Zhu-Dong-Li, and SEE estimates for the CMS; the y-axes are the abalones' ages. In the lower panels, the x-axes are the predictors derived from Zhu-Zhu, Zhu-Dong-Li, and SEE estimates of the CVS; the y-axes are the estimated absolute residuals.

We denote sample correlations for the 500 bootstrap samples as $\rho_1, \dots, \rho_{500}$. We then compute $1 - \sum_{i=1}^{500} |\rho_i|/500$ and call it the bootstrap error of the estimator. The result is summarized in the following table.

Table 5: bootstrap error of the estimators

Functionals	Estimators		
$E(Y X)$	RMAVE	Zhu-Dong-Li	SEE
	0.145	0.009	0.003
$\text{var}(Y X)$	Zhu-Zhu	Zhu-Dong-Li	SEE
	0.213	0.131	0.105

We see that SEE is the top performer for estimating both CMS and CVS, followed by the estimator of Zhu, Dong and Li (2012), and then by RMAVE. We also observe that the estimation of central variance subspace is in general less accurate than that

of central mean subspace, as has been observed in many other cases, for example in Zhu, Dong and Li (2012).

12 Discussions

In this paper we introduce a general paradigm for sufficient dimension reduction with respect to a conditional statistical functional, along with semiparametrically efficient procedures to estimate the sufficient predictors of that functional. This method is particularly useful when we want to select sufficient predictors with some specific purposes in mind, such as estimating the conditional quantiles in a population. This work is a continuation, synthesis, and refinement of previous works on nonparametric mean regression, nonparametric quantile regression, and nonparametric estimation of heteroscedasticity, under the unifying framework of SDR. It provides us with tools to explore the detailed structures of the central subspace, making SDR more specific to our goals. Our work has also substantially broadened the scope of the semiparametric approach recently introduced to SDR by Ma and Zhu (2012, 2013a, and 2013b).

In a wide range of simulation studies the SEE is shown to outperform several previously proposed estimators for conditional mean, conditional quantile, and conditional variance. Moreover, the theoretical semiparametric lower bound is approximately achieved by the actual error based on simulation. Finally, the algorithm we developed for SEE has a special advantage over that proposed in Ma and Zhu (2013a): it does not rely on any specific parameterization of the central subspace, which means we do not need to subjectively assign any element of β to be nonzero from the outset.

Acknowledgements

The authors would like to thank two referees and an associate editor for their insightful and constructive comments and suggestions, which led to significant improvement of this work. Bing Li's research is supported in part by a National Science Foundation grant (DMS-1106815). Xiangrong Yin's research is supported in part by a National Science Foundation grant (DMS-1205546).

References

1. Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. Baltimore and London.
2. Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *1994 Proceedings of the Section*

on *Physical and Engineering Sciences*, Alexandria, VA: American Statistical Association, 18–25.

3. Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983 - 992.
4. Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: *Wiley*.
5. Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Annals of Statistics*, **30**, 455–474.
6. Cook, R. D. and Weisberg, S. (1991). Discussion of “Sliced inverse regression for dimension reduction”. *Journal of the American Statistical Association*, **86**, 316–342.
7. Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, **97**, 279–294.
8. Edelman, A, Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*. **20**, 303–353.
9. Godambe, V. P. (1960). An optimum property of maximum likelihood estimation. *Annals of Mathematical Statistics*. **31**, 1208–1211.
10. Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal Smoothing in Single-Index Models. *Annals of Statistics*, **21**(1),157-178.
11. Kong, E. and Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, **28**, 730–768.
12. McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall/CRC.
13. Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Annals of Statistics*, **39**, 3182–3210.
14. Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Annals of Statistics*, **37**, 1272–1298.
15. Li, B. and McCullagh, P. (1994). Potential Functions and Conservative Estimating Functions *Annals of Statistics*. **22**, 340–356.
16. Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction, *Annals of Statistics*, **33**, 1580-1616.

17. Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316–342.
18. Li, K. -C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *Journal of the American Statistical Association*, **86**, 316–342.
19. Li, L., Li, B., and Zhu, L.-X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association*. **105**, 1188–1201.
20. Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*. **107**, 168–179.
21. Ma, Y. and Zhu, L. (2013a). Efficient estimation in sufficient dimension reduction. *Annals of Statistic*, **41**, 250–268.
22. Ma, Y. and Zhu, L. (2013b). Efficiency loss caused by linearity condition in dimension reduction. *Biometrika*, **100**, 371–383.
23. Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of American Statistical Association*, **103**, 811-821, 2008.
24. Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**, 1236–1265.
25. van der Vaart (1998). *Asymptotic Statistics*. Cambridge Press.
26. Xia, Y. Tong, H., and Li, W. K. (2002). Single-index volatility models and estimation. *Statistica Sinica*, **12**, 785–799.
27. Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363–3410.
28. Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Annals of Statistics*, **35**, 2654-2690.
29. Ye, Z. and Weiss, R.E. (2003). Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods. *Journal of the American Statistical Association*, **98**, 968–979.
30. Yin, X. and Cook, R. D. (2002). Dimension reduction for conditional k th moment in regression. *Journal of the Royal Statistical Society: Series B*, **64**, 159–175.

31. Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Annals of Statistics*, **39**, 3392–3416.
32. Yin, X., Li, B. and Cook, R. D. (2008) Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, **99**, 1733–175.
33. Zhu, L., Dong, Y., and Li, R. (2012). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statistica Sinica*. In press.
34. Zhu, L. and Zhu, L.-X. (2009). Dimension reduction for conditional variance in regression. *Statistica Sinica*, **19**, 869–883.
35. Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, **101**, 1638–1651.
36. Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, **36**, 1108–1126.