Journal of Multivariate Analysis 101 (2010) 271–290

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

An integral transform method for estimating the central mean and central subspaces

Peng Zeng [a,*], Yu Zhu [b]

[a] *Department of Mathematics and Statistics, Auburn University, Auburn, AL, 36849, USA*
[b] *Department of Statistics, Purdue University, West Lafayette, IN, 47907, USA*

ARTICLE INFO

Article history:
Received 3 June 2008
Available online 19 August 2009

AMS 2000 subject classifications:
62G08
62G20

Keywords:
Average derivative estimate
Generalized multiple index model
Integral transform
Kernel density estimation
Sufficient dimension reduction

ABSTRACT

The central mean and central subspaces of generalized multiple index model are the main inference targets of sufficient dimension reduction in regression. In this article, we propose an integral transform (ITM) method for estimating these two subspaces. Applying the ITM method, estimates are derived, separately, for two scenarios: (i) No distributional assumptions are imposed on the predictors, and (ii) the predictors are assumed to follow an elliptically contoured distribution. These estimates are shown to be asymptotically normal with the usual root-n convergence rate. The ITM method is different from other existing methods in that it avoids estimation of the unknown link function between the response and the predictors and it does not rely on distributional assumptions of the predictors under scenario (i) mentioned above.

1. Introduction

Suppose that $Y \in \mathbb{R}$ is a univariate response and $\mathbf{X} \in \mathbb{R}^p$ is a vector of continuous explanatory variables. A general model can be postulated about the relationship between Y and $\mathbf{X}$ as follows.

$$Y = g(\beta_1^\tau \mathbf{X}, \ldots, \beta_q^\tau \mathbf{X}, \varepsilon) = g(\mathcal{B}^\tau \mathbf{X}, \varepsilon), \tag{1}$$

where $\beta_1, \ldots, \beta_q$ are p-dimensional vectors ($q < p$), $\mathcal{B} = (\beta_1, \ldots, \beta_q)$, $g(\cdot)$ is an unspecified $(q+1)$-variate link function, ε is a random error independent of $\mathbf{X}$, and $E(\varepsilon) = 0$. This model was originally proposed in [1] to facilitate sufficient dimension reduction and includes many well-known models as special cases. When ε is additive, that is,

$$g(\mathcal{B}^\tau \mathbf{X}, \varepsilon) = h(\mathcal{B}^\tau \mathbf{X}) + \varepsilon, \tag{2}$$

where h is an unknown link function, (1) is known as the multiple index model in the literature. Furthermore, if $q = 1$, then the multiple index model becomes the single index model. We refer to (1) as the *generalized multiple index model*. There exists one drawback with the formulation of (1). In some applications, it may be difficult to conceive a link function or a meaningful independent random error. To avoid this drawback, [2] proposed the following model, using conditional independence,

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathcal{B}^\tau \mathbf{X}, \tag{3}$$

[*] Corresponding author.
E-mail addresses: zengpen@auburn.edu (P. Zeng), yuzhu@stat.purdue.edu (Y. Zhu).

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.08.004

where $\perp\!\!\!\perp$ means "independent of". The model (3) states that given $\mathcal{B}^\tau\mathbf{X}$, Y and $\mathbf{X}$ are independent of each other. In other words, all the information in $\mathbf{X}$ about Y is contained in the low-dimensional projection $\mathcal{B}^\tau\mathbf{X}$. Using conditional distribution functions, (3) can be re-written as

$$F_{Y|\mathbf{X}}(y|\mathbf{x}) = F_{Y|\mathcal{B}^\tau\mathbf{X}}(y|\mathcal{B}^\tau\mathbf{x}), \tag{4}$$

where $F_{Y|\mathcal{B}^\tau\mathbf{X}}(y|\mathcal{B}^\tau\mathbf{x})$ is the conditional distribution function of Y given $\mathcal{B}^\tau\mathbf{X} = \mathcal{B}^\tau\mathbf{x}$. (4) was mentioned but not explicitly explored in both [1] and [2]. Although the models (1), (3) and (4) are different in their formulations, they are in fact equivalent to each other. We state this equivalence as a lemma below and defer its proof to the Appendices.

Lemma 1. *Models (1), (3) and (4) are equivalent.*

The column vectors $\beta_1, \ldots, \beta_q$ of $\mathcal{B}$ are referred to as indices. In the literature on sufficient dimension reduction, these indices are interpreted as the directions along which Y and the projection of $\mathbf{X}$ are dependent. The linear space spanned by these indices, denoted by $\mathcal{S}(\mathcal{B})$, is referred to as a dimension reduction subspace. For a given generalized multiple index model, its dimension reduction subspace may not be unique. [2] introduced a concept, called *central subspace*, to resolve this ambiguity. The central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is defined to be the intersection of all the dimension reduction subspaces when $\mathcal{S}_{Y|\mathbf{X}}$ is a dimension reduction subspace itself. [2] also showed that the central subspace exists under general conditions. We assume its existence throughout this article. Under (1), it is not difficult to see that the mean response $E(Y|\mathbf{X} = \mathbf{x})$ also depends on a low-dimensional projection of $\mathbf{x}$. The corresponding projection space is referred to as a mean dimension reduction subspace in the literature. The intersection of all the mean dimension reduction subspaces, when it is a mean dimension reduction subspace itself, is called the *central mean subspace*, denoted by $\mathcal{S}_{E(Y|\mathbf{X})}$ ([3]). Note that the central mean subspace is always a subspace of the central subspace, that is, $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

A number of methods exist for estimating the central subspace, such as sliced inverse regression (SIR; [1]), sliced average variance estimate (SAVE; [4]), and contour regression (CR; [5]). For estimating the central mean subspace, the existing methods include the structure adaptive method (SAM; [6]), the minimum average variance estimation method (MAVE; [7]), the principal Hessian direction method (pHd; [8]), and the iterative Hessian transformation method (IHT; [3]). Among these methods, SIR, SAVE, CR, pHd and IHT avoid estimating the unknown link function, which is considered an advantage in high-dimensional scenarios, but they need to impose restrictive distributional assumptions of $\mathbf{X}$. On the other hand, the methods SAM, MAVE and their variants avoid distributional assumptions on $\mathbf{X}$, but they need to estimate the unknown link function nonparametrically. Relaxing distributional assumption on $\mathbf{X}$ and mitigating the burden of nonparametric function fitting become two major motivations in the literature for developing computationally efficient methods for estimating the central and central mean subspaces.

Under the single index model, $\mathcal{B} = (\beta_1)$, and the central and central mean subspaces are identical and equal to $\mathcal{S}(\beta_1)$. [9] proposed to estimate β_1 by averaging the derivative (or gradient) of $E(Y|\mathbf{X} = \mathbf{x})$ with respect to $\mathbf{x}$ and called the resulting estimate an average derivative estimate (ADE). The ADE method for estimating β_1 has two major advantages. First, it avoids the estimation of the link function; and second, it achieves the root-n convergence rate. See [10–12] for more discussions. The main drawback of ADE is that it can only recover one direction.

Lately, under the assumption that $\mathbf{X}$ follows the multivariate normal distribution, [13] proposed a Fourier method for exhaustively estimating the central mean and central subspaces. The Fourier method can be viewed as an extension of the ADE method, but the distributional assumption on $\mathbf{X}$ is too restrictive in some applications. In the current article, we further generalize the Fourier method in the following two directions. First, we remove the distributional assumption, and use a plug-in estimate of the log density when estimating the central and central mean subspaces. Second, the Fourier transform is extended to general integral transforms, which provide users with the flexibility of choosing the transform most suitable for their applications. We call the resulting method the integral transform (ITM) method. The ITM method does not impose distributional assumptions on $\mathbf{X}$ and avoid nonparametric fitting of the unknown link function. It does require the use of a nonparametric plug-in estimate of the log density of $\mathbf{X}$. It is known that nonparametric estimates (e.g. the plug-in estimate of log density) typically converge at a rate slower than root-n. Therefore, it is nontrivial to establish the root-n convergence rate of the estimates of the central and central mean subspaces generated by the ITM method. Following [9], we show that these estimates can still converge at the usual root-n rate when the technique "undersmoothing" is implemented. Furthermore, we derive the asymptotic distributions of the subspaces estimates explicitly. The ITM method is a full-fledged generalization of the ADE method from the single index model to the generalized multiple index model.

The rest of the article is organized as follows. Section 2 proposes the ITM method and derives matrices whose column spaces are identical to the central mean and central subspaces. These matrices are referred to as *candidate matrices* or kernel matrices in the literature. Section 3 derives the estimates of the central mean and central subspaces based on the candidate matrices, and proves their asymptotic normalities. Section 4 derives the estimates of the central mean and central subspaces when $\mathbf{X}$ follows an elliptically contoured distribution. Section 5 discusses issues related to the implementation of the proposed methods. Section 6 presents several simulation examples and Section 7 concludes this article with additional remarks.

Throughout this article we assume that all the involved distributions admit densities, and denote the joint density function of Y and $\mathbf{X}$ by $f_{Y,\mathbf{X}}(y, \mathbf{x})$, the conditional density function of Y given $\mathbf{X} = \mathbf{x}$ by $f_{Y|\mathbf{X}}(y|\mathbf{x})$, and the marginal density function of $\mathbf{X}$ by $f_{\mathbf{X}}(\mathbf{x})$. The model (4) can be restated in terms of these conditional densities as

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = f_{Y|\mathcal{B}^\tau\mathbf{X}}(y|\mathcal{B}^\tau\mathbf{x}). \tag{5}$$

We further assume these density functions are differentiable with respect to their coordinates wherever necessary. For ease of reading, technical assumptions and proofs are collected in the Appendices.

2. Integral transform and candidate matrices

We first consider the central mean subspace under the multiple index model (2). Let $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. A key fact utilized by the ADE method is

$$\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}) = \mathcal{B}\,\frac{\partial h}{\partial \mathbf{u}}(\mathbf{u})\bigg|_{\mathbf{u}=\mathcal{B}^\tau \mathbf{x}} \in \mathcal{S}(\mathcal{B}),$$

that is, the derivative of $m(\mathbf{x})$ belongs to the linear space spanned by $\mathcal{B}$. This further implies that $E(\partial m(\mathbf{X})/\partial \mathbf{x}) \in \mathcal{S}(\mathcal{B})$, which leads to the average derivative estimate. The ADE method suffers from two drawbacks. First, it fails when $E[\partial m(\mathbf{X})/\partial \mathbf{x}] = 0$; for example, when $m(\mathbf{x}) = (\beta^\tau \mathbf{x})^2$ and $\mathbf{X}$ is standard normal. Second, as mentioned in the introduction, the ADE method can only generate one direction and thus is not able to estimate the central mean subspace of dimension higher than one. The first drawback can be overcome by using a proper weight function $W(\mathbf{x})$ such that $E[W(\mathbf{X})\partial m(\mathbf{X})/\partial \mathbf{x}] \neq 0$. To further overcome the second drawback, we propose to use a family of weight functions $\{W(\mathbf{x}, \mathbf{u}) : \mathbf{u} \in \mathbb{R}^p\}$ instead of a single weight function; note that $\mathbf{u}$ is the family index varying in $\mathbb{R}^p$. Consequently different weight functions can generate different vectors in the central mean subspace. By collecting the estimates of these vectors, it is possible to derive an estimate of the entire central mean subspace. Weighted average derivatives were probably first considered by [10], but they were only used to facilitate the calculation of ADEs.

Define $\xi(\mathbf{u})$ to be the expectation of $\partial m/\partial \mathbf{x}$ weighted by $W(\mathbf{x}, \mathbf{u})$,

$$\xi(\mathbf{u}) = E\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{X}) W(\mathbf{X}, \mathbf{u})\right] = \int \frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}) W(\mathbf{x}, \mathbf{u}) f_{\mathbf{X}}(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{6}$$

In fact $\xi(\mathbf{u})$ is the integral transform of the density-weighted $\partial m/\partial \mathbf{x}$, and $W(\mathbf{x}, \mathbf{u})$ is the kernel function of this transform. When $W(\cdot, \cdot)$ is chosen properly, the linear space spanned by $\{\xi(\mathbf{u}),\ \mathbf{u} \in \mathbb{R}^p\}$ is identical to that spanned by the density-weighted derivatives $\{f_{\mathbf{X}}(\mathbf{x})\partial m(\mathbf{x})/\partial \mathbf{x},\ \mathbf{x} \in \mathbb{R}^p\}$.

Definition 1. Let $\mathbf{g}$ be a vector-valued function from $\mathbb{R}^p$ to $\mathbb{R}^p$. An integral transform with kernel $W(\mathbf{x}, \mathbf{u})$ is said to be nondegenerate for $\mathbf{g}$ if

$$\operatorname{span}\left\{\int \mathbf{g}(\mathbf{x}) W(\mathbf{x}, \mathbf{u})\mathrm{d}\mathbf{x},\ \mathbf{u} \in \mathbb{R}^p\right\} = \operatorname{span}\{\mathbf{g}(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^p\}.$$

And the kernel $W(\mathbf{x}, \mathbf{u})$ is said to be a nondegenerate kernel.

Although the nondegenerate kernel is defined for a specific function in the definition above, there exist kernels that are nondegenerate for a wide range of functions. Two such examples are $W_1(\mathbf{x}, \mathbf{u}) = \exp(\imath \mathbf{u}^\tau \mathbf{x})$ and $W_2(\mathbf{x}, \mathbf{u}) = H(\mathbf{u} - \mathbf{x})$ where $H(\cdot)$ is absolutely integrable. The integral transforms with W_1 and W_2 are commonly known as the Fourier transform and the convolution transform, respectively.

Lemma 2. *Both W_1 and W_2 are nondegenerate kernels for any absolutely integrable function* $\mathbf{g}(\mathbf{x})$.

It is known that the central mean subspace can be spanned by $\partial m(\mathbf{x})/\partial \mathbf{x}$ with $\mathbf{x} \in \operatorname{supp}(\mathbf{X})$,

$$\mathcal{S}_{E(Y|\mathbf{X})} = \operatorname{span}\left\{\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}) :\ \mathbf{x} \in \operatorname{supp}(\mathbf{X})\right\} = \operatorname{span}\left\{\frac{\partial m}{\partial \mathbf{x}}(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) :\ \mathbf{x} \in \mathbb{R}^p\right\}.$$

When W is chosen to be a nondegenerate kernel for $f_{\mathbf{X}}(\mathbf{x})\partial m(\mathbf{x})/\partial \mathbf{x}$, $\mathcal{S}_{E(Y|\mathbf{X})}$ can also be spanned by $\xi(\mathbf{u})$ with $\mathbf{u} \in \mathbb{R}^p$. Notice that in (6), $\partial m/\partial \mathbf{x}$ is assumed known. In what follows, we derive a different expression for $\xi(\mathbf{u})$ that does not require this assumption. Under some regularity conditions (see Lemma 3), applying integration by parts to (6), we have

$$\xi(\mathbf{u}) = E\left[\frac{\partial m}{\partial \mathbf{x}}(\mathbf{X}) W(\mathbf{X}, \mathbf{u})\right] = -E[Y\psi(\mathbf{X}, \mathbf{u})], \tag{7}$$

where $\psi(\mathbf{x}, \mathbf{u}) = \partial W(\mathbf{x}, \mathbf{u})/\partial \mathbf{x} + W(\mathbf{x}, \mathbf{u})\mathbf{g}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x}) = \partial \log f_{\mathbf{X}}(\mathbf{x})/\partial \mathbf{x}$ is the derivative of the log density of $\mathbf{X}$. Because the second expression of (7) does not involve m or its derivative, $\xi(\mathbf{u})$ can be calculated without involving the link function m or its derivative. We define a candidate matrix, denoted by $\mathcal{M}_{\text{ITM}}$, for $\mathcal{S}_{E(Y|\mathbf{X})}$ as follows.

$$\mathcal{M}_{\text{ITM}} = \int \xi(\mathbf{u})\xi(\mathbf{u})^\tau\,\mathrm{d}\mathbf{u} = E[\mathcal{U}_{\text{ITM}}(\mathbf{Z}_1, \mathbf{Z}_2)], \tag{8}$$

where $\mathbf{z} = (y, \mathbf{x})$, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent and identically distributed as $\mathbf{Z}$,

$$\mathcal{U}_{\text{ITM}}(\mathbf{z}_1, \mathbf{z}_2) = y_1 y_2 \int \psi(\mathbf{x}_1, \mathbf{u})\psi(\mathbf{x}_2, \mathbf{u})^\tau\,\mathrm{d}\mathbf{u}.$$

The next lemma claims that the column space of $\mathcal{M}_{\text{ITM}}$ is exactly equal to the central mean subspace $\mathcal{S}_{E(Y|\mathbf{X})}$. Thus it is indeed a candidate matrix for $\mathcal{S}_{E(Y|\mathbf{X})}$. A function is said to vanish on the boundary of supp($\mathbf{X}$) if it goes to zero when $\mathbf{x}$ goes to any point on the boundary or goes to infinity when the support of $\mathbf{X}$ is unbounded.

Lemma 3. *Assume $f_{\mathbf{X}}(\mathbf{x})\partial m(\mathbf{x})/\partial \mathbf{x}$ exists and is absolutely integrable. If $W(\mathbf{x}, \mathbf{u})m(\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ vanishes on the boundary of supp($\mathbf{X}$) and $\log f_{\mathbf{X}}(\mathbf{x})$ is differentiable, then (7) holds. Furthermore, if $W(\cdot, \cdot)$ is a nondegenerate kernel for $f_{\mathbf{X}}(\mathbf{x})\partial m(\mathbf{x})/\partial \mathbf{x}$ and $\xi(\mathbf{u})$ is square integrable, then $\mathcal{M}_{\text{ITM}}$ is a nonnegative definite matrix and $\mathcal{S}(\mathcal{M}_{\text{ITM}}) = \mathcal{S}_{E(Y|\mathbf{X})}$.*

Remark 1. The choice of $W(\cdot, \cdot)$ determines how the derivative $\partial m(\mathbf{x})/\partial \mathbf{x}$ is integrated together to give $\xi(\mathbf{u})$, and it further determines the properties of $\mathcal{M}_{\text{ITM}}$. For example, [13] chose $W(\mathbf{x}, \mathbf{u}) = \exp(\imath \mathbf{u}^\tau \mathbf{x})$ to develop the Fourier method for sufficient dimension reduction. Another possible choice is $W(\mathbf{x}, \mathbf{u}) = \delta(\mathbf{u} - \mathbf{x})$ where $\delta(\cdot)$ is the Dirac delta function. With this choice, $\xi(\mathbf{u}) = \partial m(\mathbf{u})/\partial \mathbf{x} \cdot f_{\mathbf{X}}(\mathbf{u})$, and $\mathcal{M}_{\text{ITM}}$ becomes the density-weighted outer product of $\partial m(\mathbf{u})/\partial \mathbf{x}$. The simple outer product of $\partial m(\mathbf{u})/\partial \mathbf{x}$ was discussed in [14] and [7].

Next we apply the same approach to derive a candidate matrix for the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. As discussed in the introduction, $\mathcal{S}_{E(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$. Let $T(Y)$ be an arbitrary transformation of the response Y. The central mean subspace for the transformed response $T(Y)$, denoted by $\mathcal{S}_{E[T(Y)|\mathbf{X}]}$, is defined in a similar way as $\mathcal{S}_{E(Y|X)}$. It is not difficult to see that $\mathcal{S}_{E[T(Y)|\mathbf{X}]}$ also belongs to $\mathcal{S}_{Y|\mathbf{X}}$ and is not necessarily identical to $\mathcal{S}_{E(Y|\mathbf{X})}$. This suggests that the collection of a number of different central mean subspaces (e.g., generated from different transformations) may cover the entire central subspace; furthermore, the combination of the candidate matrices of these central mean subspaces may lead to a candidate matrix of the central subspace.

We propose to use a simple family of transformations (indexed by $v \in \mathbb{R}$), that is, $\{T_v(\cdot) : T_v(y) = H(y, v),\ \text{for } y, v \in \mathbb{R}\}$, where H is a given function. Under some mild conditions, $\mathcal{S}_{Y|\mathbf{X}} = \sum_{v\in\mathbb{R}} \mathcal{S}_{E[T_v(Y)|\mathbf{X}]}$, which can be implied from Lemma 4 below. For $v \in \mathbb{R}$, the mean response of $T_v(Y)$ is $m(\mathbf{x}, v) = E[H(Y, v)|\mathbf{X} = \mathbf{x}]$. Similar to the definition of $\xi(\mathbf{u})$ for the central mean subspace $\mathcal{S}_{E(Y|X)}$, we define $\xi(\mathbf{u}, v)$ as the integral transform of $\partial m(\mathbf{x}, v)/\partial \mathbf{x} f_{\mathbf{X}}(\mathbf{x})$ using a kernel $W(\mathbf{x}, \mathbf{u})$. Under certain regularity conditions (given in Lemma 4),

$$\xi(\mathbf{u}, v) = E\left[\frac{\partial m(\mathbf{X}, v)}{\partial \mathbf{x}} W(\mathbf{X}, \mathbf{u})\right] = -E[H(Y, v)\psi(\mathbf{X}, \mathbf{u})]. \tag{9}$$

The second equality above is obtained by integration by parts.

The vectors $\xi(\mathbf{u}, v)$ for $\mathbf{u} \in \mathbb{R}^p$ and $v \in \mathbb{R}$ play the same role for $\mathcal{S}_{Y|\mathbf{X}}$ as $\xi(\mathbf{u})$ for $\mathcal{S}_{E(Y|\mathbf{X})}$. They span the central subspace and can be used to form a candidate matrix for $\mathcal{S}_{Y|\mathbf{X}}$. We define

$$\mathcal{M}_{\text{ITC}} = \iint \xi(\mathbf{u}, v)\xi(\mathbf{u}, v)^\tau \, d\mathbf{u} dv = E[\mathcal{U}_{\text{ITC}}(\mathbf{Z}_1, \mathbf{Z}_2)], \tag{10}$$

where

$$\mathcal{U}_{\text{ITC}}(\mathbf{z}_1, \mathbf{z}_2) = \int H(y_1, v)H(y_2, v)\, dv \int \psi(\mathbf{x}_1, \mathbf{u})\psi(\mathbf{x}_2, \mathbf{u})^\tau \, d\mathbf{u}.$$

Notice that the only difference between (8) and (10) is that Y_1Y_2 is used in $\mathcal{M}_{\text{ITM}}$ whereas $\int H(Y_1, v)H(Y_2, v)dv$ is used in $\mathcal{M}_{\text{ITC}}$. The next lemma states that $\mathcal{M}_{\text{ITC}}$ is indeed a candidate matrix for $\mathcal{S}_{Y|\mathbf{X}}$.

Lemma 4. *Assume $\partial f_{Y|\mathbf{X}}(y|\mathbf{x})/\partial \mathbf{x} \cdot f_{\mathbf{X}}(\mathbf{x})$ exists and is absolutely integrable. If for any given y and $\mathbf{u}$, $W(\mathbf{x}, \mathbf{u})f_{Y,\mathbf{X}}(y, \mathbf{x})$ vanishes on the boundary of supp($\mathbf{X}$), then (9) holds. Furthermore, if $H(y, v)W(\mathbf{x}, \mathbf{u})$ is a nondegenerate kernel for $f_{\mathbf{X}}(\mathbf{x})\partial f_{Y|\mathbf{X}}(y|\mathbf{x})/\partial \mathbf{x}$ and $\xi(\mathbf{u}, v)$ is square integrable, then $\mathcal{M}_{\text{ITC}}$ is a nonnegative definite matrix and $\mathcal{S}(\mathcal{M}_{\text{ITC}}) = \mathcal{S}_{Y|\mathbf{X}}$.*

The candidate matrix $\mathcal{M}_{\text{ITC}}$ is naturally connected with the candidate matrix $\mathcal{M}_{\text{SIR}}$ (=cov[$E(\mathbf{X}|Y)$]) used in SIR. When $\mathbf{X}$ follows the p-dimensional standard normal distribution, $\mathbf{g}(\mathbf{x}) = -\mathbf{x}$. If we further choose $W(\mathbf{x}, \mathbf{u}) \equiv 1$, then it can be verified that

$$\mathcal{M}_{\text{ITC}} = E\left[\int H(Y_1, v)H(Y_2, v)\, dv\ E(\mathbf{X}_1|Y_1)E(\mathbf{X}_2|Y_2)^\tau\right].$$

Based on the expression above, if we select $H(\cdot, \cdot)$ such that $\int H(Y_1, v)H(Y_2, v)\, dv$ is positive on the support of Y, then

$$\mathcal{S}(\mathcal{M}_{\text{ITC}}) = \text{span}\{E(\mathbf{X}|Y = y),\ y \in \text{supp}(Y)\} = \text{span}(\mathcal{M}_{\text{SIR}}).$$

It is known that SIR fails to capture the directions along which the link function g is even and the distribution of $\mathbf{X}$ is symmetric. Hence, when the weight function $W(\cdot, \cdot)$ is chosen to be a constant function, $\mathcal{M}_{\text{ITC}}$ is degenerated to be equivalent to $\mathcal{M}_{\text{SIR}}$ and may also fail to recover the whole central subspace. In general as claimed in Lemma 4, $\mathcal{M}_{\text{ITC}}$ can successfully recover the whole central subspace $\mathcal{S}_{Y|\mathbf{X}}$ when $W(\mathbf{x}, \mathbf{u})$ is chosen appropriately.

3. Estimates of candidate matrices

Let $\mathbf{Z}_i = (Y_i, \mathbf{X}_i)$ for $1 \le i \le n$ be n iid copies of $(Y, \mathbf{X})$ and $\mathbf{z}_i = (y_i, \mathbf{x}_i)$ be a realization of $\mathbf{Z}_i$ for $1 \le i \le n$. In what follows, we first consider the derivation of an estimate for $\mathcal{M}_{\text{ITC}}$. Note that $\mathcal{M}_{\text{ITC}} = E[\mathcal{U}_{\text{ITC}}(\mathbf{Z}_1, \mathbf{Z}_2)]$. Since $\mathcal{M}_{\text{ITC}}$ is the expectation of $\mathcal{U}_{\text{ITC}}(\mathbf{Z}_1, \mathbf{Z}_2)$, a natural estimate of $\mathcal{M}_{\text{ITC}}$ is the sample average of $\mathcal{U}_{\text{ITC}}(\mathbf{z}_i, \mathbf{z}_j)$. There however remains one difficulty, which is $\mathbf{g}(\mathbf{x}) = \partial \log f_{\mathbf{X}}(\mathbf{x})/\partial \mathbf{x} = (\frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}))/f_{\mathbf{X}}(\mathbf{x})$ is unknown. Therefore, we need to estimate $\mathbf{g}(\mathbf{x})$ first based on $\{\mathbf{x}_i\}_{1\le i\le n}$, an iid sample from $f_{\mathbf{X}}(\mathbf{x})$.

If $f_{\mathbf{X}}(\mathbf{x})$ can be assumed to belong to a parametric family, that is, $f_{\mathbf{X}}(\mathbf{x}) = f_0(\mathbf{x}; \theta)$ where f_0 is of known form and θ is a vector of unknown parameters, then $\mathbf{g}(\mathbf{x})$ can be estimated parametrically. In this article, our focus is on estimating the central mean and central subspaces without imposing distributional assumptions on $\mathbf{X}$. Hence, we do not assume any parametric form for $f_{\mathbf{X}}(\mathbf{x})$ and instead propose to estimate $f_{\mathbf{X}}(\mathbf{x})$ and its derivative nonparametrically. The kernel density estimate of $f_{\mathbf{X}}$ at a fixed point $\mathbf{x}_0$ is

$$\widehat{f_{\mathbf{X}}(\mathbf{x}_0)} = \hat{f}_h(\mathbf{x}_0) = (nh^p)^{-1} \sum_{\ell=1}^{n} K\left(\frac{\mathbf{x}_0 - \mathbf{x}_\ell}{h}\right),$$

where $K(\cdot)$ is a kernel function and h is the bandwidth [15,16]. Note that h depends on n, but we use h instead of h_n for the cleanness of expression. The derivative of $f_{\mathbf{X}}(\mathbf{x}_0)$ is estimated by the derivative of $\hat{f}_h(\mathbf{x}_0)$. Both $\hat{f}_h(\mathbf{x}_0)$ and $\partial\hat{f}_h(\mathbf{x}_0)/\partial\mathbf{x}$ are asymptotically consistent estimates of $f_{\mathbf{X}}(\mathbf{x}_0)$ and $\frac{\partial}{\partial \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}_0)$, and they lead to an asymptotically consistent estimate of $\mathbf{g}(\mathbf{x}_0)$,

$$\hat{\mathbf{g}}(\mathbf{x}_0) = \frac{\frac{\partial}{\partial \mathbf{x}}\hat{f}_h(\mathbf{x}_0)}{\hat{f}_h(\mathbf{x}_0)} = \frac{(nh^{p+1})^{-1} \sum_{\ell=1}^{n} K'((\mathbf{x}_0 - \mathbf{x}_\ell)/h)}{(nh^{p})^{-1} \sum_{\ell=1}^{n} K((\mathbf{x}_0 - \mathbf{x}_\ell)/h)} \tag{11}$$

where $K'(\cdot)$ is the derivative of $K(\cdot)$. Given $\{\mathbf{z}_i\}_{1\le i\le n}$ and $\hat{\mathbf{g}}(\cdot)$, we derive an estimate of $\mathcal{M}_{\text{ITC}}$,

$$\hat{\mathcal{M}}^*_{\text{ITCk}} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\mathcal{U}}_{\text{ITCk}}(\mathbf{z}_i, \mathbf{z}_j),$$

where $\hat{\mathcal{U}}_{\text{ITCk}}$ is obtained by replacing $\mathbf{g}$ in $\mathcal{U}_{\text{ITC}}$ by $\hat{\mathbf{g}}$.

Let $\hat{I}_i = I_{[\hat{f}_h(\mathbf{x}_i) > b_n]}$ for $1 \le i \le n$ where I is an indicator function and b_n is a pre-specified threshold. To avoid the negative effect of small values of $\hat{f}_h$, we modified $\hat{\mathcal{M}}^*_{\text{ITCk}}$ to be

$$\hat{\mathcal{M}}_{\text{ITCk}} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\mathcal{U}}_{\text{ITCk}}(\mathbf{z}_i, \mathbf{z}_j)\hat{I}_i\hat{I}_j.$$

To ensure that $\hat{\mathcal{M}}_{\text{ITCk}}$ is asymptotically consistent, b_n needs to decrease to zero as the sample size n goes to infinity. In the rest of the article, we suppress the subscript of b_n and simply use b. The next theorem states that under some technical conditions $\hat{\mathcal{M}}_{\text{ITCk}}$ is asymptotically normal with a worked out covariance matrix.

Theorem 2. *Suppose conditions* (A1)–(A4) *and* (A5c)–(A9c) *hold. If* (a) $n \to \infty$, $h \to 0$, $b \to 0$, *and* $b^{-1}h \to 0$; (b) *for some* $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^{2p+2} \to \infty$; *and* (c) $nh^{2s-2} \to 0$, *then* $vec(\hat{\mathcal{M}}_{\text{ITCk}})$ *asymptotically follows a multivariate normal distribution,*

$$\sqrt{n}\,(vec(\hat{\mathcal{M}}_{\text{ITCk}}) - vec(\mathcal{M}_{\text{ITC}})) \xrightarrow{\mathcal{L}} N(0, \mathbf{\Sigma}_{\text{ITCk}}),$$

where $\mathbf{\Sigma}_{\text{ITCk}}$ *is the covariance matrix of* $vec(\mathcal{R}(\mathbf{Z}) + \mathcal{R}(\mathbf{Z})^\tau)$, *and*

$$\mathcal{R}(\mathbf{z}) = \iint \xi(\mathbf{u}, v)\left\{[m(\mathbf{x}, v) - H(y, v)]\psi(\mathbf{x}, \mathbf{u}) + \frac{\partial m(\mathbf{x}, v)}{\partial \mathbf{x}} W(\mathbf{x}, \mathbf{u})\right\}^\tau \mathrm{d}\mathbf{u}\mathrm{d}v.$$

The operator vec$(\cdot)$ in the above theorem is to convert a matrix to a vector by stacking up all its columns. For example, if $\mathcal{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_k)$ is a $p \times k$ matrix with columns $\mathbf{m}_1, \ldots, \mathbf{m}_k$, vec$(\mathcal{M}) = (\mathbf{m}_1^\tau, \ldots, \mathbf{m}_k^\tau)^\tau$ is a pk-dimensional vector. The proof of Theorem 2 given in Appendix A.4 is an extension of the proof of Theorem 3.1 in [9].

The estimation of $\mathcal{M}_{\text{ITC}}$ (or equivalently $\mathcal{S}_{Y|\mathbf{X}}$) is a semiparametric estimation problem. As a matter of fact, $\mathcal{M}_{\text{ITC}}$ can be considered a finite-dimensional parameter and (10) can be regarded as a mapping from the density function $f_{Y,\mathbf{X}}(y, \mathbf{x}) = f_{Y|\mathbf{X}}(y|\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ to the space of $p \times p$ semi-positive definite matrices. The proposed integral transform method has successfully transformed the original semiparametric model to an estimation problem of $(\mathcal{M}_{\text{ITC}}, f_{\mathbf{X}}(\mathbf{x}))$ with $f_{\mathbf{X}}(\mathbf{x})$ being a nuisance, infinite-dimensional parameter. Note that the estimation of the other nuisance parameter $f_{Y|\mathbf{X}}$ has been avoided. The standard approach to estimating a finite-dimensional parameter (e.g., $\mathcal{M}_{\text{ITC}}$) at the presence of an infinite-dimensional parameter (e.g., $f_{\mathbf{X}}(\mathbf{x})$) is to employ a plug-in estimate of the latter ([17]). In the derivation of $\hat{\mathcal{M}}_{\text{ITCk}}$, a kernel estimate of $f_{\mathbf{X}}(\mathbf{x})$ is used as the plug-in estimate. Typically the nonparametric plug-in estimate cannot achieve the root-n convergence rate, whereas the estimate of the finite parameter can often attain the rate. Theorem 2 shows it is indeed the case for $\hat{\mathcal{M}}_{\text{ITCk}}$. The root-$n$ convergence rate of $\hat{\mathcal{M}}_{\text{ITCk}}$ is achieved through a technique called "undersmoothing" originally used by [9]. The condition (c) in Theorem 2 requires the bandwidth h be narrower than the usual optimal bandwidth for kernel density estimation. A narrower bandwidth makes the bias of $\hat{f}_h(\mathbf{x}_i)$ vanish at a rate faster than $\sqrt{n}$; it however results in larger variabilities to $\hat{f}_h(\mathbf{x}_i)$'s. Because $\hat{f}_h(\mathbf{x}_i)$'s are averaged in $\hat{\xi}(\mathbf{u}, v)$, the fast decrease in bias is inherited by $\hat{\xi}(\mathbf{u}, v)$ and the increased variabilities are mitigated. Therefore, the root-n convergence rate of $\hat{\mathcal{M}}_{\text{ITCk}}$ is obtained.

Theorem 2 implies that the eigenvalues and eigenvectors of $\hat{\mathcal{M}}_{\text{ITCk}}$ also converge to those of $\mathcal{M}_{\text{ITC}}$ at the same convergence rate. If the rank of $\mathcal{M}_{\text{ITC}}$ is q, the space spanned by the eigenvectors corresponding to the largest q eigenvalues of $\hat{\mathcal{M}}_{\text{ITCk}}$ converges to $\mathcal{S}(\mathcal{M}_{\text{ITC}})$, or equivalently $\mathcal{S}_{Y|\mathbf{X}}$, at the root-$n$ rate.

The foregoing procedure can be modified to derive an estimate of $\mathcal{M}_{\text{ITM}}$. Because the modification is straightforward, it is omitted and only the results are reported below. The estimate of $\mathcal{M}_{\text{ITM}}$ is

$$\hat{\mathcal{M}}_{\text{ITMk}} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\mathcal{U}}_{\text{ITMk}}(\mathbf{z}_i, \mathbf{z}_j) \hat{I}_i \hat{I}_j$$

where $\hat{\mathcal{U}}_{\text{ITMk}}$ is obtained by replacing $\mathbf{g}$ in $\mathcal{U}_{\text{ITM}}$ by $\hat{\mathbf{g}}$ and $\hat{I}_i$ for $1 \le i \le n$ are the same as before. The asymptotic behavior of $\hat{\mathcal{M}}_{\text{ITMk}}$ is described in the following theorem.

Theorem 3. *Suppose conditions* (A1)–(A4) *and* (A5m)–(A9m) *hold. If* (a) $n \to \infty, h \to 0, b \to 0,$ *and* $b^{-1}h \to 0$; (b) *for some* $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^{2p+2} \to \infty$; *and* (c) $nh^{2s-2} \to 0$, *then* $vec(\hat{\mathcal{M}}_{\text{ITMk}})$ *asymptotically follows a multivariate normal distribution,*

$$\sqrt{n}\,(vec(\hat{\mathcal{M}}_{\text{ITMk}}) - vec(\mathcal{M}_{\text{ITM}})) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{\Sigma}_{\text{ITMk}}),$$

where $\boldsymbol{\Sigma}_{\text{ITMk}}$ *is the covariance matrix of* $vec(\mathcal{R}(\mathbf{Z}) + \mathcal{R}(\mathbf{Z})^\tau)$, *and*

$$\mathcal{R}(\mathbf{z}) = \int \xi(\mathbf{u}) \left\{ [m(\mathbf{x}) - y]\psi(\mathbf{x}, \mathbf{u}) + \frac{\partial m(\mathbf{x})}{\partial \mathbf{x}} W(\mathbf{x}, \mathbf{u}) \right\}^\tau \, d\mathbf{u}.$$

The proof of Theorem 3 is similar to that of Theorem 2; see Remark 3 in Appendix A.4. Theorem 3 asserts that $\hat{\mathcal{M}}_{\text{ITMk}}$ converges to $\mathcal{M}_{\text{ITM}}$ at the root-n rate, which implies that the eigenvalues and eigenvectors of $\hat{\mathcal{M}}_{\text{ITMk}}$ also converge to those of $\mathcal{M}_{\text{ITM}}$ at the same rate. If the rank of $\mathcal{M}_{\text{ITM}}$ is q, the space spanned by the eigenvectors corresponding to the largest q eigenvalues of $\hat{\mathcal{M}}_{\text{ITMk}}$ converges to $\mathcal{S}(\mathcal{M}_{\text{ITM}})$, or equivalently $\mathcal{S}_{E(Y|\mathbf{X})}$, at the same root-$n$ rate.

If we choose $W(\mathbf{x}, \mathbf{u}) \equiv 1$, then $\mathcal{M}_{\text{ITM}} = \boldsymbol{\delta}\boldsymbol{\delta}^\tau$ and $\hat{\mathcal{M}}_{\text{ITM}} = \hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^\tau$ where $\boldsymbol{\delta} = E[\partial m(\mathbf{X})/\partial \mathbf{x}]$ is the average derivative and $\hat{\boldsymbol{\delta}}$ is the ADE of $\boldsymbol{\delta}$ proposed in [9]. According to Theorem 3.1 in [9], the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})$ is equal to the covariance matrix of $\mathbf{r}(\mathbf{X}, Y) = [m(\mathbf{X}) - Y]\mathbf{g}(\mathbf{X}) + \partial m(\mathbf{X})/\partial \mathbf{x}$. It can be verified that the asymptotic covariance of $\sqrt{n}(\hat{\boldsymbol{\delta}}\hat{\boldsymbol{\delta}}^\tau - \boldsymbol{\delta}\boldsymbol{\delta}^\tau)$ is equal to that of $\boldsymbol{\delta}\mathbf{r}(\mathbf{X}, Y)^\tau + \mathbf{r}(\mathbf{X}, Y)\boldsymbol{\delta}^\tau$, which is exactly equal to $\boldsymbol{\Sigma}_{\text{ITMk}}$ in Theorem 3. Therefore, Theorem 3 is a successful generalization of Theorem 3.1 in [9].

4. Elliptically contoured distributions

The kernel estimate $\hat{f}_{\mathbf{X}}$ used in both $\hat{\mathcal{M}}_{\text{ITCk}}$ and $\hat{\mathcal{M}}_{\text{ITMk}}$ is susceptible to the curse of dimensionality; when the dimension of $\mathbf{X}$ increases, its performance deteriorates quickly. The performance of $\hat{\mathcal{M}}_{\text{ITCk}}$ and $\hat{\mathcal{M}}_{\text{ITMk}}$, however, does not degrade as fast as $\hat{f}_{\mathbf{X}}$, especially when the dimension of $\mathbf{X}$ is in teens. When the dimension of $\mathbf{X}$ is above 20, the performance of $\hat{\mathcal{M}}_{\text{ITCk}}$ and $\hat{\mathcal{M}}_{\text{ITMk}}$ starts to become unacceptable. When some prior knowledge is available regarding the distribution of $\mathbf{X}$, estimates that are less susceptible to the curse of dimensionality can be developed. In this section, we consider the family of elliptically contoured distributions, which is broad enough to include many important multivariate distributions such as multivariate normal distributions as its members. Elliptically contoured distributions are essentially one-dimensional because their density functions are radial functions. The contour regression (CR) method proposed by [5] for estimating the central subspace also assumes that $\mathbf{X}$ follows an elliptically contoured distribution.

Since an elliptically contoured distribution can always be transformed to become spherical, we only focus on spherical distributions below. Let $f_{\mathbf{X}}(\mathbf{x}) = f(\mathbf{x}^\tau \mathbf{x})$ with mean 0 and covariance $\mathcal{I}_p$. Let $R = \|\mathbf{X}\|$ and $f_R(r)$ be the density function of R. Then $f_R(r)$ can be represented in terms of $f_{\mathbf{X}}(\mathbf{x})$,

$$f_R(r) = (2\pi^{p/2}) \left(\Gamma\left(\frac{1}{2}p\right)\right)^{-1} r^{p-1} f(r^2),$$

where $\Gamma(\cdot)$ is the Gamma function; see [18]. This relationship provides the possibility to estimate the derivative of $\log f(\mathbf{x}^\tau \mathbf{x})$ with respect to $\mathbf{x}$, i.e., $\mathbf{g}(\mathbf{x})$, through estimating the derivative of $\log f_R(r)$ with respect to r. In fact

$$\mathbf{g}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \log f(\mathbf{x}^\tau \mathbf{x}) = \frac{2\mathbf{x} f'(\mathbf{x}^\tau \mathbf{x})}{f(\mathbf{x}^\tau \mathbf{x})} = \frac{\mathbf{x}}{r} g_R(r) - \frac{p-1}{r^2}\mathbf{x},$$

where $g_R(r) = f_R'(r)/f_R(r)$, and $f'(\cdot)$ and $f_R'(\cdot)$ are the derivatives of $f(\cdot)$ and $f_R(\cdot)$, respectively.

Recall that $\{\mathbf{x}_i\}_{i=1}^n$ is an iid sample from $f_{\mathbf{X}}(\mathbf{x})$. Let $r_i = \|\mathbf{x}_i\|$. Then $\{r_i\}_{i=1}^n$ is an iid sample from $f_R(r)$. Let $\tilde{f}_{R,h}$ be the kernel density estimator based on $\{r_i\}_{i=1}^n$ with bandwidth h, and let $\tilde{f}'_{R,h}$ be the derive of $\tilde{f}_{R,h}$. Then an estimate of $\mathbf{g}(\mathbf{x})$ is

$$\tilde{\mathbf{g}}(\mathbf{x}_i) = \frac{\mathbf{x}_i}{r_i}\frac{\tilde{f}'_{R,h}(r_i)}{\tilde{f}_{R,h}(r_i)} - \frac{p-1}{r_i^2}\mathbf{x}_i.$$

Replacing $\hat{\mathbf{g}}(\mathbf{x})$ in $\hat{\mathcal{M}}_{\mathrm{ITCk}}$ by $\tilde{\mathbf{g}}(\mathbf{x})$, we obtain an estimate of $\mathcal{M}_{\mathrm{ITC}}$,

$$\tilde{\mathcal{M}}_{\mathrm{ITCe}} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \tilde{\mathcal{U}}_{\mathrm{ITCe}}(\mathbf{z}_i, \mathbf{z}_j)\tilde{I}_i\tilde{I}_j,$$

where $\tilde{\mathcal{U}}_{\mathrm{ITCe}}(\mathbf{z}_i, \mathbf{z}_j)$ is obtained by replacing $\mathbf{g}$ in $\mathcal{U}_{\mathrm{ITC}}(\mathbf{z}_i, \mathbf{z}_j)$ by $\tilde{\mathbf{g}}$, $\tilde{I}_i = I_{[\tilde{f}_{R,h}(r_i)>b]}$ is used to trim the points where the estimated densities are too small, and the subscript e in $\hat{\mathcal{M}}_{\mathrm{ITCe}}$ indicates that $\mathbf{X}$ follows an elliptically contoured distribution. The asymptotic behavior of $\hat{\mathcal{M}}_{\mathrm{ITCe}}$ is stated in the next theorem.

Theorem 4. *Suppose that* $\mathbf{X}$ *follows an elliptically contoured distribution with mean* 0 *and covariance matrix* $\mathcal{I}_p$ *and some regularity conditions hold. If* (a) $n \to \infty$, $h \to 0$, $b \to 0$, *and* $b^{-1}h \to 0$; (b) *for some* $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^4 \to \infty$; *and* (c) $nh^{2s-2} \to 0$, *then the estimate* $\hat{\mathcal{M}}_{\mathrm{ITCe}}$ *asymptotically follows a normal distribution, that is,*

$$\sqrt{n}(vec(\tilde{\mathcal{M}}_{\mathrm{ITCe}}) - vec(\mathcal{M}_{\mathrm{ITC}})) \xrightarrow{\mathcal{L}} N(0, \mathbf{\Sigma}_{\mathrm{ITCe}}),$$

where $\mathbf{\Sigma}_{\mathrm{ITCe}}$ *is the covariance matrix of* $vec(\mathcal{R}(\mathbf{Z}) + \mathcal{R}(\mathbf{Z})^\tau)$, *and*

$$\mathcal{R}(\mathbf{z}) = E[\mathcal{U}_{\mathrm{ITC}}(\mathbf{Z}_1, \mathbf{z})] + \iint \xi(\mathbf{u}, v)\left\{\mathbf{m}_1(r, \mathbf{u}, v)g_R(r) + \frac{\partial \mathbf{m}_1(r, \mathbf{u}, v)}{\partial r}\right\}^\tau d\mathbf{u}dv,$$

and $\mathbf{m}_1(r, \mathbf{u}, v) = r^{-1}E[H(Y, v)W(\mathbf{X}, \mathbf{u})\mathbf{X}|R = r]$.

Due to space limitation, the regularity conditions required by Theorem 4 are omitted; see Remark 4 in Appendix A.4 for more details. Similarly, for the central mean subspace, we replace $\hat{\mathbf{g}}(\mathbf{x})$ in $\hat{\mathcal{M}}_{\mathrm{ITMk}}$ by $\tilde{\mathbf{g}}(\mathbf{x})$ and derive the following estimate of $\mathcal{M}_{\mathrm{ITM}}$,

$$\tilde{\mathcal{M}}_{\mathrm{ITMe}} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n \tilde{\mathcal{U}}_{\mathrm{ITMe}}\tilde{I}_i\tilde{I}_j,$$

where $\tilde{\mathcal{U}}_{\mathrm{ITMe}}$ is obtained by replacing $\mathbf{g}$ in $\mathcal{U}_{\mathrm{ITM}}$ by $\tilde{\mathbf{g}}$. The following theorem claims that vec($\tilde{\mathcal{M}}_{\mathrm{ITMe}}$) asymptotically follows a normal distribution. The regularity conditions are also omitted due to space limitation; see Remark 4 in Appendix A.4.

Theorem 5. *Suppose that* $\mathbf{X}$ *follows an elliptically contoured distribution with mean* 0 *and covariance matrix* $\mathcal{I}_p$, *and some regularity conditions hold.* (a) $n \to \infty$, $h \to 0$, $b \to 0$, *and* $b^{-1}h \to 0$; (b) *for some* $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^4 \to \infty$; *and* (c) $nh^{2s-2} \to 0$, *then the estimate* $\tilde{\mathcal{M}}_{\mathrm{ITMe}}$ *asymptotically follows a normal distribution, that is,*

$$\sqrt{n}(vec(\tilde{\mathcal{M}}_{\mathrm{ITMe}}) - vec(\mathcal{M}_{\mathrm{ITM}})) \xrightarrow{\mathcal{L}} N(0, \mathbf{\Sigma}_{\mathrm{ITMe}}),$$

where $\mathbf{\Sigma}_{\mathrm{ITMe}}$ *is the covariance matrix of* $vec(\mathcal{R}(\mathbf{Z}) + \mathcal{R}(\mathbf{Z})^\tau)$, *and*

$$\mathcal{R}(\mathbf{z}) = E[\mathcal{U}_{\mathrm{ITM}}(\mathbf{Z}_1, \mathbf{z})] + \int \xi(\mathbf{u})\left\{\mathbf{m}_1(r, \mathbf{u})g_R(r) + \frac{\partial \mathbf{m}_1(r, \mathbf{u})}{\partial r}\right\}^\tau d\mathbf{u},$$

and $\mathbf{m}_1(r, \mathbf{u}) = r^{-1}E[YW(\mathbf{X}, \mathbf{u})\mathbf{X}|R = r]$.

When $\mathbf{X}$ does not follow an elliptically contoured distribution, the ITM method using $\tilde{\mathcal{M}}_{\mathrm{ITCe}}$ or $\tilde{\mathcal{M}}_{\mathrm{ITMe}}$ is not able to estimate central or central mean subspace correctly in general. However, we may be able to alleviate the departure from the elliptically contoured distribution via a reweighting scheme as suggested by [19]. [20] showed that the shapes of lower-dimensional projections from high-dimensional data are mostly elliptically contoured in some sense. This suggests that the elliptically contoured distribution may be a reasonable distribution to work with in analyzing high-dimensional data.

5. Implementation

To implement the methods proposed in the previous sections, the weight functions $W(\cdot,\cdot)$ and $H(\cdot,\cdot)$ need to be specified. Gaussian kernels are popular choices for weight functions. Let

$$W(\mathbf{x},\mathbf{u}) = (2\pi\sigma_u^2)^{-p/2}\ \exp\{-(\mathbf{x}-\mathbf{u})^\tau(\mathbf{x}-\mathbf{u})/(2\sigma_u^2)\};$$
$$H(y,v) = (2\pi\sigma_v^2)^{-1/2}\ \exp\{-(y-v)^2/(2\sigma_v^2)\}.$$

Simple calculation yields

$$\int \psi(\mathbf{x}_1,\mathbf{u})\psi(\mathbf{x}_2,\mathbf{u})\mathrm{d}\mathbf{u} = c_1 \exp\left\{-\frac{\mathbf{x}_{12}^\tau\mathbf{x}_{12}}{4\sigma_u^2}\right\}\left(\frac{1}{2\sigma_u^2}\mathcal{I}_p + \left(\mathbf{g}(\mathbf{x}_1) - \frac{1}{2\sigma_u^2}\mathbf{x}_{12}\right)\left(\mathbf{g}(\mathbf{x}_2) + \frac{1}{2\sigma_u^2}\mathbf{x}_{12}\right)^\tau\right),$$
$$\int H(y_1,v)H(y_2,v)\mathrm{d}v = c_2\exp\left\{-\frac{(y_1-y_2)^2}{4\sigma_v^2}\right\},$$

where $\mathbf{x}_{12} = \mathbf{x}_1 - \mathbf{x}_2$ and c_1 and c_2 are two positive constants only depending on σ_u and σ_v. With the specified weight functions W and H and the worked out integrals above, explicit expressions of $\hat{\mathcal{M}}_{\mathrm{ITMk}}$, $\hat{\mathcal{M}}_{\mathrm{ITCk}}$, $\tilde{\mathcal{M}}_{\mathrm{ITMe}}$ and $\tilde{\mathcal{M}}_{\mathrm{ITCe}}$ are available. Due to space limitation, these expressions are omitted. Gaussian kernels are also used in $\hat{\mathbf{g}}(\mathbf{x})$ and $\tilde{\mathbf{g}}(\mathbf{x})$. The choice of bandwidths for these kernels will be discussed later.

Next we present a unified algorithm for estimating the central mean and central subspaces using various estimated candidate matrices. For ease of presentation, we use $\mathcal{S}$ to denote a subspace, which can be a central mean subspace or a central subspace; use $\mathcal{M}$ to denote a candidate matrix, which can be $\mathcal{M}_{\mathrm{ITC}}$ or $\mathcal{M}_{\mathrm{ITM}}$; and use $\hat{\mathcal{M}}$ to denote an estimated candidate matrix, which can be $\hat{\mathcal{M}}_{\mathrm{ITCk}}$, $\tilde{\mathcal{M}}_{\mathrm{ITCe}}$, $\hat{\mathcal{M}}_{\mathrm{ITMk}}$, or $\tilde{\mathcal{M}}_{\mathrm{ITMe}}$. Let $\hat{\mathcal{S}}$ denote an estimate of $\mathcal{S}$. Suppose the dimension of $\mathcal{S}$ is known to be q and all the involved tuning parameters are already given. The unified algorithm for computing $\hat{\mathcal{S}}$ consists of the five steps given below.

(1) Specify the parameters q, σ_u^2, σ_v^2, h, and b.

(2) Calculate $\tilde{\mathbf{x}}_i = \hat{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ and $\tilde{y}_i = (y_i - \bar{y})/s_y$, where $\bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}$ are the sample mean and covariance matrix of $\mathbf{x}_i$'s, and $\bar{y}$ and s_y are the sample mean and standard deviation of y_i's.

(3) Calculate $\hat{\mathcal{M}}$ using the standardized data $\{(\tilde{y}_i, \tilde{\mathbf{x}}_i)\}_{1\le i\le n}$.

(4) Perform the spectral decomposition of $\hat{\mathcal{M}}$ and obtain the eigenvalue–eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$ with $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$.

(5) $\mathcal{S}$ is estimated by $\hat{\mathcal{S}} = \mathrm{span}\{\hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\mathbf{e}}_1, \ldots, \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\mathbf{e}}_q\}$.

In application, the specifications of $\mathcal{M}$ and its estimate in the above algorithm depend on which subspace is of interest and whether $\mathbf{X}$ follows an elliptically contoured distribution. For example, if one is interested in the central subspace and $\mathbf{X}$ can be assumed to follow an elliptically contoured distribution, then $\mathcal{M}_{\mathrm{ITC}}$ and $\tilde{\mathcal{M}}_{\mathrm{ITCe}}$ should be used in the places of $\mathcal{M}$ and $\hat{\mathcal{M}}$, respectively. We want to mention that, even when the central mean subspace is of interest, it is still worth to check the central subspace, because there may exist some unsuspected, important patterns.

Before proceeding to discuss other issues in implementation, we introduce a measure of distance between two linear subspaces. This distance measure will be used to measure the discrepancy between $\mathcal{S}$ and $\hat{\mathcal{S}}$. Suppose $\mathcal{A}$ and $\mathcal{B}$ are two $p\times q$ matrices of full column rank, and $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$ are their column spaces. Let $\mathcal{P}_{\mathcal{A}} = \mathcal{A}(\mathcal{A}^\tau\mathcal{A})^{-1}\mathcal{A}^\tau$ and $\mathcal{P}_{\mathcal{B}} = \mathcal{B}(\mathcal{B}^\tau\mathcal{B})^{-1}\mathcal{B}^\tau$ be the projection matrices onto $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$, respectively. Let $r = \sqrt{\mathrm{tr}(\mathcal{P}_{\mathcal{A}}\mathcal{P}_{\mathcal{B}})/q}$. It can be shown that r is always in $[0, 1]$, and we refer to r as the trace correlation between $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$. The larger r is, the closer $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$ are to each other. Hence, we define the distance between $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$ to be $D(\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B})) = 1 - r$. It can be shown that $D(\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B})) = 0$ if $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$ are identical; and $D(\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B})) = 1$ if $\mathcal{S}(\mathcal{A})$ and $\mathcal{S}(\mathcal{B})$ are orthogonal to each other; otherwise, $D(\mathcal{S}(\mathcal{A}), \mathcal{S}(\mathcal{B}))$ is strictly between 0 and 1. The discrepancy between $\mathcal{S}$ and $\hat{\mathcal{S}}$ is measured by $D(\hat{\mathcal{S}}, \mathcal{S})$.

5.1. Choice of tuning parameters

There are four tuning parameters (σ_u^2, σ_v^2, h and b) that need to be specified in Step 1. The parameters σ_u^2 and σ_v^2 are associated with the Gaussian weight functions W and H, h is the bandwidth of the Gaussian kernel function used in $\hat{\mathbf{g}}(\mathbf{x})$ or $\tilde{\mathbf{g}}(\mathbf{x})$, and b is the threshold used in trimming. The parameter σ_v^2 is not needed when estimating the central mean subspace. Based on experience from extensive simulation study, we recommend to choose b to trim 10% of the data points, and choose $\sigma_u^2 = 5.0$ and $\sigma_v^2 = 0.5$ for samples of moderate size. In general, a bootstrap procedure can be engineered to select σ_u^2, σ_v^2 and h, following the original idea of [21] and its implementation in [13]. The key idea is to select the parameters to minimize the variability of $D(\hat{\mathcal{S}}, \mathcal{S})$. The bootstrap procedure is to estimate this variability by the average distance between $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}}^{(j)}$, where $\hat{\mathcal{S}}^{(j)}$ is an estimate of $\mathcal{S}$ based on a bootstrapped sample. In what follows, we use the selection of σ_u^2 as an illustrative example. Suppose we want to select the optimal σ_u^2 from a set of candidate values $\{\sigma_1^2, \ldots, \sigma_m^2\}$, which are in general equally spaced in an interval. For each ℓ such that $1 \le \ell \le m$,

(1) randomly sample from $\{(y_i, \mathbf{x}_i)\}_{1\le i\le n}$ with replacement to generate N bootstrapped samples each of size n, and the jth sample is denoted by $\{(y_i^{(j)}, \mathbf{x}_i^{(j)})\}_{1\le i\le n}$ for $1 \le j \le N$;
(2) for each bootstrapped sample, e.g., the jth sample $\{(y_i^{(j)}, \mathbf{x}_i^{(j)})\}_{1\le i\le n}$, derive the estimate of $\mathcal{S}$ and denote it by $\hat{\mathcal{S}}^{(j)}$;
(3) calculate the distance between $\hat{\mathcal{S}}^{(j)}$ and $\hat{\mathcal{S}}$ and denote it by $d^{(j)} = D(\hat{\mathcal{S}}^{(j)}, \hat{\mathcal{S}})$;
(4) calculate $\bar{d}(\sigma_\ell^2) = N^{-1}\sum_{j=1}^{N} d^{(j)}$, which is the average distance between $\hat{\mathcal{S}}^{(j)}$ and $\hat{\mathcal{S}}$ for $1 \le j \le N$.

The optimal σ_u^2 is chosen to be the σ_ℓ^2 that minimizes $\bar{d}(\sigma_\ell^2)$. The above procedure can be easily modified for the selection of σ_v^2 or h.

Sometimes, it is necessary to choose all the tuning parameters. Two possible approaches can be followed. The first is to select the parameters iteratively, for example, to choose h and b first, then σ_u^2, and at last σ_v^2. The second approach is to choose the optimal tuning parameters from a set of candidate 4-tuples $\{(\sigma_{ui}^2, \sigma_{vi}^2, h_i, b_i)\}_{1\le i\le m}$. The latter approach is often more intensive computationally.

5.2. Selection of dimensionality

In many applications, the dimension of the central or central mean subspace is unknown and needs to be inferred from data. One simple approach to determining the dimension is to plot the ordered eigenvalues of $\hat{\mathcal{M}}$ as in principal components analysis and look for an "elbow" pattern in the plot. The dimension is chosen to be the number of dominant eigenvalues. Although subjective, this method is intuitive and works well in general. Alternatively, the bootstrap procedure described in the previous subsection can be adopted to determine the dimension of $\mathcal{S}$. The procedure treats the dimension of $\mathcal{S}$ as another tuning parameter and then choose the dimension k that minimizes $\bar{d}(k)$. More details can be found in both [21] and [13]. It is also possible to conduct a formal hypothesis test to determine the dimension of $\mathcal{S}$. The hypotheses $H_0 : \dim(\mathcal{S}) = d$ versus $H_1 : \dim(\mathcal{S}) > d$ is equivalent to $H_0 : \mathrm{rank}(\mathcal{M}) = d$ versus $H_1 : \mathrm{rank}(\mathcal{M}) > d$, because $\mathcal{S}(\mathcal{M}) = \mathcal{S}$. A proper test statistic for the latter is $\hat{\Lambda}_d = n\sum_{i=p}^{d+1}\hat{\lambda}_i$, where $\hat{\lambda}_{d+1}, \dots, \hat{\lambda}_p$ are the smallest $(p-d)$ eigenvalues of $\hat{\mathcal{M}}$. Under H_0, the smallest $(p-d)$ eigenvalues of $\hat{\mathcal{M}}$ are zeros, and thus $\hat{\lambda}_{d+1}, \dots, \hat{\lambda}_p$ are expected to be small. Therefore we reject H_0 when $\hat{\Lambda}_d$ is large. We conjecture that $\hat{\Lambda}_d$ follows a weighted chi-squared distribution asymptotically. This conjecture is currently under investigation. For practical use, we may utilize a permutation algorithm discussed in [22] to evaluate the P-value of this test.

6. Simulation examples

We use ITCk, ITCe, ITMk and ITMe to label the methods for estimating $S_{Y|\mathbf{X}}$ or $S_{E(Y|\mathbf{X})}$ based on the estimated candidate matrices $\hat{\mathcal{M}}_{\text{ITCk}}$, $\tilde{\mathcal{M}}_{\text{ITCe}}$, $\hat{\mathcal{M}}_{\text{ITMk}}$ and $\tilde{\mathcal{M}}_{\text{ITMe}}$, respectively. Let $\mathbf{1}_n$ denote an n-dimensional vector of ones and $\mathbf{0}_n$ denote an n-dimensional vector of zeros.

Example 1. Consider the following model,

$$Y = \frac{|\beta_1^\tau \mathbf{X} + \varepsilon|}{2 + |\beta_2^\tau \mathbf{X} - 4 + \varepsilon|},$$

where $\beta_1^\tau = (\mathbf{1}_4^\tau, \mathbf{0}_6^\tau)$, $\beta_2^\tau = (\mathbf{0}_6^\tau, \mathbf{1}_4^\tau)$, $\mathbf{X}$ follows a mixture of three multivariate normal distributions, and ε is an error term following the standard normal distribution and independent of $\mathbf{X}$. The distribution of $\mathbf{X}$ is $0.5N(\mathbf{c}_1, \mathcal{I}_{10}) + 0.3N(-\mathbf{c}_1, \mathcal{I}_{10}) + 0.2N(\mathbf{c}_2, \mathcal{I}_{10})$, where $\mathbf{c}_1^\tau = (1, -1, 1, -1, 1, -1, 1, -1, 1, -1)$, and $\mathbf{c}_2^\tau = (1, 1, -1, -1, -1, -1, -1, -1, 1, 1)$. We randomly draw 500 samples (each of size 1000) from the above model, and apply ITCk ($\sigma_u^2 = 5.0, \sigma_v^2 = 0.5$), ITMk ($\sigma_u^2 = 5.0$), SIR and SAVE to each sample to estimate the central subspace. Assume the dimension of the subspace is known to be 2. In order to compare the methods under different specifications of bandwidth and/or the number of slices (i.e., H), we run ITCk and ITMk with $h = 0.5, 1.0, 1.5, 2.0$ and SIR and SAVE with $H = 4, 6, 8, 10$, separately. In total, 16 estimated central subspaces are obtained for each sample. The distances (i.e., $D(\hat{\mathcal{S}}, \mathcal{S})$) between these estimated central subspaces and the true central subspace $\mathcal{S} = \mathrm{span}\{\beta_1, \beta_2\}$ are calculated. Thus for each method under each specified h or H, 500 such distances are generated; and these distances are used to generate the boxplots in Fig. 1. The mean and standard deviations of these distances are reported in Table 1. Overall, ITMk demonstrates the best performance in this example. The boxplots for ITCk shows that its performance is sensitive to the choice of bandwidth; and the best choice of bandwidth for this example is $h = 1.5$. Both ITCk and ITMk outperform SIR and SAVE. The reason the latter two methods do not perform as well is that the linearity condition they require is violated by the mixture distribution of $\mathbf{X}$.

Example 2. Consider the following model,

$$Y = \beta_1^\tau \mathbf{X} + (4\beta_2^\tau \mathbf{X})\varepsilon,$$

where $\beta_1^\tau = (\mathbf{1}_4^\tau, \mathbf{0}_6^\tau)$, $\beta_2^\tau = (\mathbf{0}_6^\tau, \mathbf{1}_4^\tau)$, $\mathbf{X}$ follows the same mixture distribution as in Example 1, and ε follows $N(0, 1)$ and independent of $\mathbf{X}$. Similarly as in Example 1, we randomly draw 500 samples of size 1000 from the above model, and apply

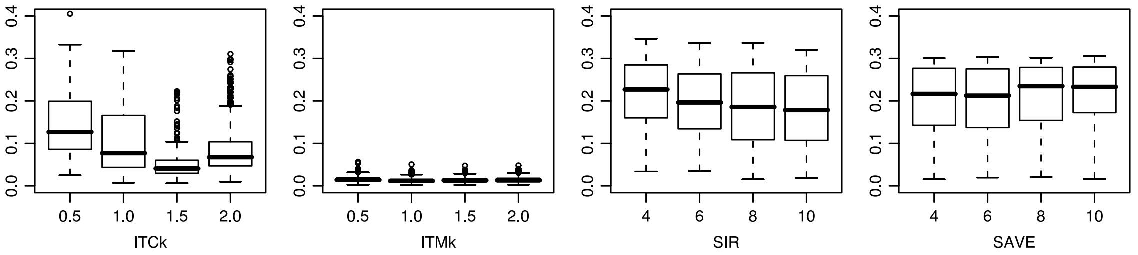

Fig. 1. Boxplots of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods with different values of h or H in Example 1. For ITCk and ITMk, the four boxplots correspond to $h = 0.5$, 1.0, 1.5, 2.0; and for SIR and SAVE, the four boxplots correspond to $H = 4, 6, 8, 10$.

Table 1
Means and standard deviations of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods with different values of h or H in Example 1.

	$h = 0.5$	$h = 1.0$	$h = 1.5$	$h = 2.0$
ITCk	0.1468 (0.0792)	0.1127 (0.0859)	0.0490 (0.0313)	0.0833 (0.0548)
ITMk	0.0156 (0.0074)	0.0125 (0.0060)	0.0139 (0.0064)	0.0144 (0.0067)
	$H = 4$	$H = 6$	$H = 8$	$H = 10$
SIR	0.2178 (0.0711)	0.1953 (0.0766)	0.1843 (0.0843)	0.1818 (0.0843)
SAVE	0.2012 (0.0787)	0.2010 (0.0785)	0.2122 (0.0749)	0.2182 (0.0677)

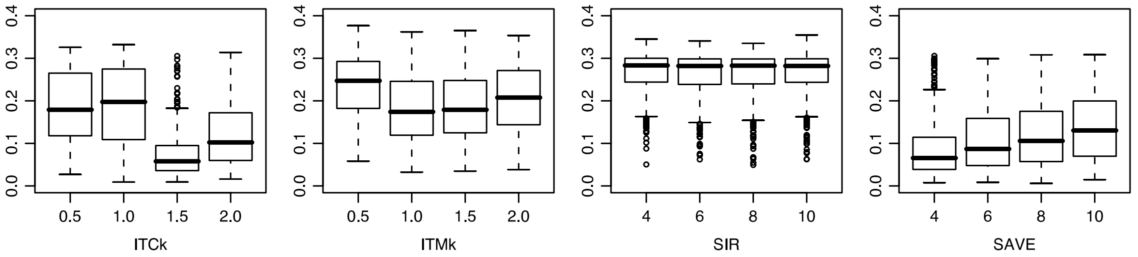

Fig. 2. Boxplots of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods with different values of h and H in Example 2. For ITCk and ITMk, the four boxplots correspond to $h =$ 0.5, 1.0, 1.5, 2.0, respectively; and for SIR and SAVE, the four boxplots corresponds to $H = 4, 6, 8, 10$.

Table 2
Means and standard deviations of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods with different values of h or H in Example 2.

	$h = 0.5$	$h = 1.0$	$h = 1.5$	$h = 2.0$
ITCk	0.1847 (0.0828)	0.1859 (0.0902)	0.0732 (0.0518)	0.1229 (0.0769)
ITMk	0.2356 (0.0674)	0.1832 (0.0773)	0.1866 (0.0762)	0.2060 (0.0745)
	$H = 4$	$H = 6$	$H = 8$	$H = 10$
SIR	0.2666 (0.0476)	0.2631 (0.0493)	0.2618 (0.0523)	0.2631 (0.0517)
SAVE	0.0891 (0.0708)	0.1114 (0.0803)	0.1239 (0.0802)	0.1422 (0.0825)

ITCk ($\sigma_u^2 = 5.0, \sigma_v^2 = 0.5$), ITMk ($\sigma_u^2 = 5.0$), SIR and SAVE to each sample to estimate the central subspace. Assume the dimension of the subspace is known to be 2. Again, different values of h (or H) are considered for ITCk and ITMk (or SIR and SAVE). The boxplots of the distances between estimated central subspaces and the true central subspace are generated for each method with each specification of h or H (Fig. 2) and the means and standard deviations of the distances are given in Table 2. Overall, ITCk with $h = 1.5$ demonstrates the best performance, and the second best performance belongs to SAVE with $H = 4$. ITMk is expected to have poor performance because it targets the central mean subspace spanned by β_1 only.

Example 3. Consider the following model

$$Y = e^{X_1} + (X_2 + 1.5)^2 + \varepsilon,$$

where $X_3, \ldots, X_{10}, \varepsilon$ are independent $N(0, 1)$ random variables, and X_1 and X_2 are generated from $X_3, \ldots, X_8$ and two additional independent $N(0, 1)$ random variables δ_1 and δ_2 by

$$\begin{aligned} X_1 &= 0.2X_3 + 0.2(X_4 + 2)^2 + 0.2X_5 + \delta_1, \\ X_2 &= 0.1 + 0.1(X_6 + X_7) + 0.3(X_7 + 1.5)^2 + 0.3X_8 + \delta_2. \end{aligned}$$

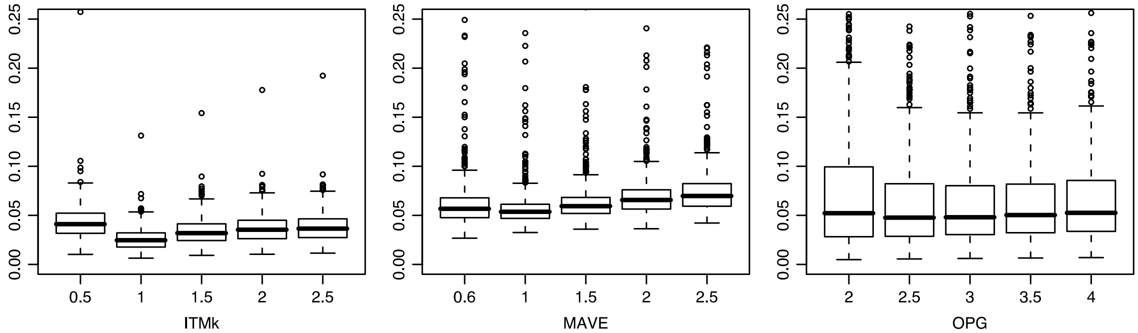

Fig. 3. Boxplots of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods with different values of bandwidth h in Example 3.

Table 3
Means and standard deviations of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods in Example 3.

ITMk	$h = 0.5$	$h = 1.0$	$h = 1.5$	$h = 2.0$	$h = 2.5$
Mean	0.0438	0.0261	0.0344	0.0372	0.0385
Standard deviation	0.0183	0.0117	0.0140	0.0150	0.0156
MAVE	$h = 0.6$	$h = 1.0$	$h = 1.5$	$h = 2.0$	$h = 2.5$
Mean	0.0648	0.0591	0.0639	0.0702	0.0751
Standard deviation	0.0363	0.0269	0.0234	0.0248	0.0264
OPG	$h = 2.0$	$h = 2.5$	$h = 3.0$	$h = 3.5$	$h = 4.0$
Mean	0.0747	0.0653	0.0632	0.0644	0.0664
Standard deviation	0.0635	0.0536	0.0485	0.0471	0.0473

Thus there exists a nonlinear confounding among the predictors. The central mean subspace is spanned by $(1, \mathbf{0}_9^\tau)^\tau$ and $(0, 1, \mathbf{0}_8^\tau)^\tau$. We use this model to compare the performance of ITMk with OPG and MAVE ([7]). The dimension of the central subspace is assumed to be known. Randomly draw 500 samples of size 1000 from this model, and apply ITMk ($\sigma_u^2 = 5.0$), OPG and MAVE to each sample to estimate the central mean subspace. Fig. 3 displays the side-by-side boxplots of $D(\hat{\mathcal{S}}, \mathcal{S})$ for ITMk, MAVE and OPG with different bandwidths, and their corresponding means and variances are reported in Table 3.

Because the initial estimates used in MAVE are OPG estimates, MAVE in general performs better than OPG. We observe that although all the methods estimate the central mean subspace with high accuracy, ITMk performs better than the other two methods in this example.

Example 4. Consider the following model,

$$Y = \frac{X_1}{0.5 + (1.5 + X_2)^2} + (1 + X_2)^2 + 0.8\,\varepsilon,$$

where $\mathbf{X} = (X_1, \ldots, X_{10})$ follows the multivariate t-distribution with 10 degrees of freedom, and ε follows $N(0, 1)$ and independent of $\mathbf{X}$. In this model, the central subspace and the central mean subspace are identical, and both are spanned by $(1, \mathbf{0}_9^\tau)^\tau$ and $(0, 1, \mathbf{0}_8^\tau)^\tau$. We use this example to compare the performance of different methods under the assumption that $\mathbf{X}$ follows an elliptically contoured distribution. In particular, we want to compare the proposed methods with SCR and GCR, which are the simple and general contour regression methods proposed in [5].

We draw 500 random samples each of size 1000 from the above model, apply SCR, GCR, SIR, SAVE, pHd, ITCe ($\sigma_u^2 = 5.0$, $\sigma_v^2 = 0.5$, and $h = 0.3$), and ITMe ($\sigma_u^2 = 5.0$ and $h = 0.3$) to each sample, and calculate $D(\hat{\mathcal{S}}, \mathcal{S})$, the distance between the true subspace and an estimated subspace. For each method, we construct a boxplot of the distances to show its overall performance (Fig. 4), and include the means and standard deviations of these distances in Table 4. All the methods except pHd have performed well in estimating $\mathcal{S}$, and ITMe has performed slightly better than the others. This is expected because SCR, GCR, ITMc and ITMe are specifically developed for handling elliptically contoured $\mathbf{X}$. The good performance of SIR is due to the fact that the linearity condition holds for the multivariate t distribution.

7. Conclusion

In this article, we have proposed the ITM method for estimating the central mean and central subspaces under the generalized multiple index model. The asymptotic properties of the resulting estimates have been established. The two major advantages of this approach are that (i) it avoids the estimation of the link function between the response and

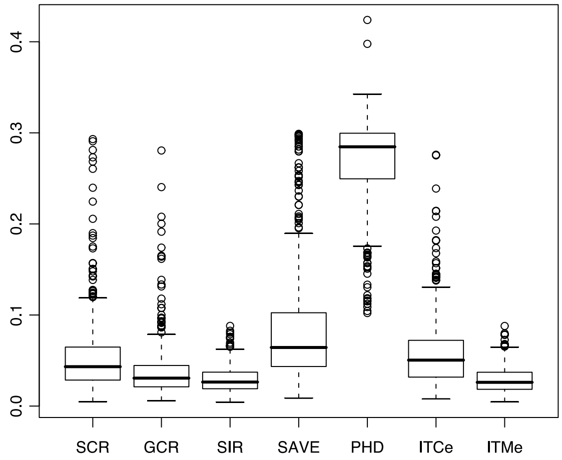

Fig. 4. Boxplots of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods in Example 4. Every boxplot is based on 500 samples of size 1000. For ITCe, $\sigma_u^2 = 5.0, \sigma_v^2 = 0.5$, and $h = 0.3$. For ITMe, $\sigma_u^2 = 5.0$ and $h = 0.3$.

Table 4
Means and standard deviations of $D(\hat{\mathcal{S}}, \mathcal{S})$ for different methods in Example 4.

	SCR	GCR	SIR	SAVE	PHD	ITCE	ITME
Mean	0.0538	0.0381	0.0295	0.0863	0.2696	0.0583	0.0288
Standard deviation	0.0422	0.0305	0.0143	0.0650	0.0456	0.0385	0.0141

predictors and (ii) it does not impose distributional assumptions on the predictors. The ITM approach is fairly flexible and can be easily extended to other regression settings such as those involving multiple responses and/or categorical responses. Due to their generality, the proposed methods may not perform well when the number of predictors is very large. When applied to specific applications, however, they can be modified and thus improved by adopting a proper integral transform and using a plug-in estimate of the marginal distribution of the predictors less susceptible to the curse of dimensionality. We will further study these issues in the future.

Acknowledgments

We thank the Associate Editor and an anonymous referee for their helpful comments and suggestions. We also thank Professor Bing Li for sharing computer codes with us. The work of Peng Zeng was supported in part by NSF grant DMS-0706880, and the work of Yu Zhu was supported in part by NSF grant DMS-0707004.

Appendix

As a convention, plain capital letters such as Y represent random variables; bold capital letters such as $\mathbf{Z}$ and $\mathbf{X}$ represent random vectors; and small letters (e.g. $\mathbf{z}$, y, and $\mathbf{x}$) represent realizations of random variables or vectors (e.g. $\mathbf{Z}$, Y and $\mathbf{X}$, respectively). Expectations such as $E[b_1(\mathbf{Z}_1, \mathbf{Z}_2)]$ and $E[b_1(\mathbf{z}_i, \mathbf{Z}_2)]$ are understood to be taken over the random variables or vectors.

A.1. Proofs in Sections 1 and 2

Proof of Lemma 1. The equivalence between (3) and (4) directly follows from the definition of conditional independence ([23]). Hence, it is enough to show that (1) and (3) are equivalent.

First, we assume that (1) holds. Given $\mathcal{B}^\tau \mathbf{X} = \mathcal{B}^\tau \mathbf{x}$, Y depends on ε only. Because ε and $\mathbf{X}$ are independent of each other, Y is independent of $\mathbf{X}$. Therefore (3) holds.

Second, we assume (3) holds. Because (3) and (4) are equivalent, we have

$$F_{Y|\mathbf{X}}(y|\mathbf{x}) = P(Y \le y|\mathbf{X} = \mathbf{x}) = P(Y \le y|\mathcal{B}^\tau \mathbf{X} = \mathcal{B}^\tau \mathbf{x}) = F_{Y|\mathcal{B}^\tau \mathbf{X}}(y|\mathcal{B}^\tau \mathbf{x}).$$

Introduce a random variable ε, which follows uniform distribution $U(0, 1)$ and is independent of $\mathbf{X}$. For any given $\mathbf{x}$, define a new random variable $\tilde{Y}$ by $\tilde{Y} = F_{Y|\mathbf{X}}^{-1}(\varepsilon|\mathbf{x})$, where $F_{Y|\mathbf{X}}^{-1}(\cdot|\mathbf{x})$ is defined to be

$$F_{Y|\mathbf{X}}^{-1}(t|\mathbf{x}) = \inf\{y : F_{Y|\mathbf{X}}(y|\mathbf{x}) \ge t\} = \inf\{y : F_{Y|\mathcal{B}^\tau \mathbf{X}}(y|\mathcal{B}^\tau \mathbf{x}) \ge t\},$$

for $0 < t < 1$. Hence, $\tilde{Y}$ is a well-defined function of $\mathcal{B}^\tau \mathbf{x}$ and ε. Denote the function by $\tilde{g}$. So we have $\tilde{Y} = \tilde{g}(\mathcal{B}^\tau \mathbf{x}, \varepsilon)$. Replacing $\mathbf{x}$ with $\mathbf{X}$, we have $\tilde{Y} = \tilde{g}(\mathcal{B}^\tau \mathbf{X}, \varepsilon)$. Clearly $\tilde{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathcal{B}^\tau \mathbf{X}$. Next we need to show that $(\tilde{Y}, \mathbf{X})$ and $(Y, \mathbf{X})$ are stochastically equivalent, and it is sufficient to show that, given $\mathbf{X} = \mathbf{x}$, $\tilde{Y}$ and Y have the same distribution.

Following the same argument as in the proof of Theorem 14.1 in [24], $F^{-1}_{Y|\mathbf{X}}(\varepsilon|\mathbf{x}) \le y$ if and only if $\varepsilon \le F_{Y|\mathbf{X}}(y|\mathbf{x})$. Then

$$\begin{aligned} P(\tilde{Y} \le u \mid \mathbf{X} = \mathbf{x}) &= P\{F^{-1}_{Y|\mathbf{X}}(\varepsilon \mid \mathbf{x}) \le u\} = P\{\varepsilon \le F_{Y|\mathbf{X}}(u|\mathbf{x})\} \\ &= F_{Y|\mathbf{X}}(u|\mathbf{x}) = P(Y \le u \mid \mathbf{X} = \mathbf{x}). \end{aligned}$$

Note that the second last equality follows from $\varepsilon \sim U(0, 1)$. Therefore, (1) holds. So the three models are equivalent. □

Proof of Lemma 2. Denote the Fourier transform of $\mathbf{g}$ as $\mathcal{F}(\mathbf{g})$. For any vector $\mathbf{b}$, it is known that $\mathbf{b}^\tau \mathbf{g} \equiv 0$ a.s. if and only if $\mathcal{F}(\mathbf{b}^\tau \mathbf{g}) = \mathbf{b}^\tau \mathcal{F}(\mathbf{g}) \equiv 0$ a.s. ([25]). So $\mathbf{g}$ and $\mathcal{F}(\mathbf{g})$ span the same linear space. Consequently, $W_1(\mathbf{x}, \mathbf{u}) = \exp(\imath \mathbf{u}^\tau \mathbf{x})$ is a nondegenerate kernel for $\mathbf{g}$.

Similarly, when $W_2(\mathbf{x}, \mathbf{u}) = H(\mathbf{u} - \mathbf{x})$, $\mathcal{F}(\int \mathbf{g}(\mathbf{x}) H(\mathbf{u} - \mathbf{x}) d\mathbf{x}) = \mathcal{F}(\mathbf{g}) \cdot \mathcal{F}(H)$. As long as H is not always zero,

$$\operatorname{span}\left\{\int \mathbf{g}(\mathbf{x}) H(\mathbf{u} - \mathbf{x}) d\mathbf{x}\right\} = \operatorname{span}\{\mathcal{F}(\mathbf{g})\} = \operatorname{span}\{\mathbf{g}\}.$$

Hence $W_2(\mathbf{x}, \mathbf{u})$ is a nondegenerate kernel for $\mathbf{g}$. □

A.2. Regularity conditions

A function $h(\mathbf{x}, \mathbf{u}, v)$ is locally Lipschitz continuous in $\mathbf{x}$ if there exists a function $\omega_h(\mathbf{x}, \mathbf{u}, v)$ such that $\|h(\mathbf{x} + \mathbf{t}, \mathbf{u}, v) - h(\mathbf{x}, \mathbf{u}, v)\| \le \omega_h(\mathbf{x}, \mathbf{u}, v)\|\mathbf{t}\|$ for $\mathbf{t}$ in a neighborhood of 0. A function $h(\mathbf{x})$ is locally γ-Hölder continuous if there exist $\gamma > 0$ and $c(\mathbf{x})$ such that $\|h(\mathbf{x} + \mathbf{v}) - h(\mathbf{x})\| \le c(\mathbf{x})\|\mathbf{v}\|^\gamma$. The functions $\omega_h(\mathbf{x}, \mathbf{u}, v)$ and $c(\mathbf{x})$ are called modulus.

(A1) The support Ω of $f_\mathbf{X}$ is a convex subset of $\mathbb{R}^p$ with nonempty interior. The underlying measure of $(y, \mathbf{x})$ can be written as $\mu_y \times \mu_\mathbf{x}$, where $\mu_\mathbf{x}$ is Lebesgue measure.
(A2) $f_\mathbf{X}(\mathbf{x})$ vanishes on the boundary of Ω.
(A3) All derivatives of $f_\mathbf{X}(\mathbf{x})$ of order s exist, where $s \ge p + 2$. Let $f^{(s)}_\mathbf{X}$ denote any sth order derivative of f. Then $f^{(s)}_\mathbf{X}$ is locally γ-Hölder continuous with modulus $c(\mathbf{x})$.
(A4) The kernel function $K(\mathbf{u})$ has support $\{\mathbf{u} \mid \|\mathbf{u}\| \le 1\}$, is symmetric, and $K(\mathbf{u}) = 0$ for all $\mathbf{u} \in \{\mathbf{u} \mid \|\mathbf{u}\| = 1\}$. The first $s - 1$ moments of $K(\mathbf{u})$ vanish. The $(s + \gamma)$-th moments of $K(\cdot)$ exist.

Given below are the additional technical conditions required by Theorem 2.

(A5c) $\int E[H(Y_1, v)|\mathbf{x}_1]E[H(Y_2, v)|\mathbf{x}_2]\mathrm{d}v \int W(\mathbf{x}_1, \mathbf{u})W(\mathbf{x}_2, \mathbf{u})\mathrm{d}\mathbf{u}$ has continuous second-order partial derivatives on $\Omega_0 \times \Omega_0 \subset \Omega \times \Omega$, where $\Omega \times \Omega - \Omega_0 \times \Omega_0$ is a set of measure 0.
(A6c) The following integrals exist.

$$\int E|H(Y, v)|^4 \mathrm{d}v, \qquad \int E|W(\mathbf{X}, \mathbf{u})|^4 \mathrm{d}\mathbf{u}, \qquad \int E\left\|\frac{\partial}{\partial \mathbf{x}} W(\mathbf{X}, \mathbf{u})\right\|^4 \mathrm{d}\mathbf{u}, \qquad \int E\|W(\mathbf{X}, \mathbf{u})\mathbf{g}(\mathbf{X})\|^4 \mathrm{d}\mathbf{u}.$$

(A7c) Denote $I_1 = I_{[f_\mathbf{X}(\mathbf{x}_1) \ge b]}$ and $\bar{I}_1 = 1 - I_1$. As $n \to \infty$, the following statements hold.

$$\int E[|H(Y_1, v)|^2 \bar{I}_1]\mathrm{d}v \int E[\|W(\mathbf{X}_1, \mathbf{u})\mathbf{g}(\mathbf{X}_1)\|^2 \bar{I}_1]\,\mathrm{d}\mathbf{u} = o(n^{-1/2}),$$

$$\int E[|H(Y_1, v)|^2 \bar{I}_1]\mathrm{d}v \int E\left[\left\|\frac{\partial W(\mathbf{X}_1, \mathbf{u})}{\partial \mathbf{x}}\right\|^2 \bar{I}_1\right] \mathrm{d}\mathbf{u} = o(n^{-1/2}).$$

(A8c) Assume that $f_\mathbf{X}, f'_\mathbf{X}, m\frac{\partial W}{\partial \mathbf{x}}, \frac{\partial m}{\partial \mathbf{x}} W$, and $mW\mathbf{g}$ are locally Lipschitz continuous in $\mathbf{x}$ with modulus $\omega_f, \omega_{f'}, \omega_{mW'}, \omega_{m'W}$, and ω_{mWg}, respectively. We further assume that $\int E\|W(\mathbf{X}, u)H^*(\mathbf{X})\|^4 \mathrm{d}\mathbf{u}$ is finite for $H^* \in \{f^{-1}_\mathbf{X} \omega_{f'}, \mathbf{g} f^{-1}_\mathbf{X} \omega_f\}$. Assume that $\iint E(\omega^* f^*)^2 \mathrm{d}\mathbf{u}\mathrm{d}v$ is finite for $\omega^* \in \{\omega_{mW'}, \omega_{m'W}, \omega_{mWg}\}$ and $f^* \in \{1, f'_\mathbf{X}, \omega_{f'}, \mathbf{g} f_\mathbf{X}, \mathbf{g}\omega_f\}$.
(A9c) Let $A_n = \{\mathbf{x} : f_\mathbf{X}(\mathbf{x}) > b\}$. The following integrals are bounded as $n \to \infty$.

$$\int\int_{A_n} |E[H(Y, v)|\mathbf{x}]|^2 |f^*(\mathbf{x})| \mathrm{d}\mathbf{x}\mathrm{d}v, \qquad \int\int_{A_n} \left\|\frac{\partial}{\partial \mathbf{x}} W(\mathbf{x}, \mathbf{u})\right\|^2 |f^*(\mathbf{x})| \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{u},$$

$$\int\int_{A_n} |W(\mathbf{x}, \mathbf{u})|^2 |f^*(\mathbf{x})| \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{u}, \qquad \int\int_{A_n} \|W(\mathbf{x}, \mathbf{u})\mathbf{g}(\mathbf{x})\|^2 |f^*(\mathbf{x})| \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{u}$$

for $f^* \in \{f_\mathbf{X}, f^{(s)}_\mathbf{X}, h^\gamma c\}$.

The technical conditions (A5m)–(A9m) required by Theorem 3 resemble (A5c)–(A9c). They can be obtained from (A5c)–(A9c) by changing $H(Y, v)$ to Y and removing $\int \mathrm{d}v$. We do not list them explicitly to save space.

A.3. A generic theorem

Considering the similarities between Theorems 2 and 3, we introduce the following notation to present them in a unified way. Let $\mathcal{M} = E[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{Z}_2)]$ and

$$\mathcal{U}_1(\mathbf{z}_1, \mathbf{z}_2) = b_1(\mathbf{z}_1, \mathbf{z}_2)\mathbf{g}(\mathbf{x}_1)\mathbf{g}(\mathbf{x}_2)^\tau + \mathbf{b}_2(\mathbf{z}_1, \mathbf{z}_2)\mathbf{g}(\mathbf{x}_2)^\tau + \mathbf{g}(\mathbf{x}_1)\mathbf{b}_2(\mathbf{z}_2, \mathbf{z}_1)^\tau + \mathcal{B}_3(\mathbf{z}_1, \mathbf{z}_2),$$

where $b_1(\mathbf{z}_1, \mathbf{z}_2)$ is a real-valued function, $\mathbf{b}_2(\mathbf{z}_1, \mathbf{z}_2)$ is a vector-valued function, and $\mathcal{B}_3(\mathbf{z}_1, \mathbf{z}_2)$ is a matrix-valued function. Given an iid sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$, an estimate of $\mathcal{M}$ is

$$\hat{\mathcal{M}}_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} [b_1(\mathbf{z}_i, \mathbf{z}_j)\hat{\mathbf{g}}(\mathbf{x}_i)\hat{\mathbf{g}}(\mathbf{x}_j)^\tau + \mathbf{b}_2(\mathbf{z}_i, \mathbf{z}_j)\hat{\mathbf{g}}(\mathbf{x}_j)^\tau + \hat{\mathbf{g}}(\mathbf{x}_i)\mathbf{b}_2(\mathbf{z}_j, \mathbf{z}_i)^\tau + \mathcal{B}_3(\mathbf{z}_i, \mathbf{z}_j)]\hat{I}_i\hat{I}_j,$$

where $\hat{\mathbf{g}}(\mathbf{x}_i)$ is defined in (11).

It can be verified that $\mathcal{M}$ and $\hat{\mathcal{M}}_n$ become $\mathcal{M}_{\text{ITC}}$ and $\hat{\mathcal{M}}_{\text{ITCk}}$ when b_1, $\mathbf{b}_2$ and $\mathcal{B}_3$ are properly chosen; see Appendix A.4. Similarly a different specification of b_1, $\mathbf{b}_2$ and $\mathcal{B}_3$ can make $\mathcal{M}$ and $\hat{\mathcal{M}}_n$ become $\mathcal{M}_{\text{ITM}}$ and $\hat{\mathcal{M}}_{\text{ITMk}}$. In what follows, we first prove a generic theorem, i.e., Theorem 6, without referring to particular specifications of b_1, $\mathbf{b}_2$ and $\mathcal{B}_3$. After Theorem 6 is proven, Theorems 2 and 3 can be proven simply by verifying that their corresponding specifications of b_1, $\mathbf{b}_2$ and $\mathcal{B}_3$ satisfy the conditions required by Theorem 6.

Under the regularity conditions required by Theorem 6, which will be given later, we can obtain an expansion of $\hat{\mathcal{M}}_n$ as follows.

$$\hat{\mathcal{M}}_n = \mathcal{M} + n^{-1} \sum_{i=1}^{n} [\mathcal{R}(\mathbf{z}_i) + \mathcal{R}(\mathbf{z}_i)^\tau - 2\mathcal{M}] + o_p(n^{-1/2}), \tag{A.1}$$

where

$$\mathcal{R}(\mathbf{z}) = E_{\mathbf{Z}_1}[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{z})] - E_{\mathbf{Z}_1}[a_1(\mathbf{X}_1, \mathbf{x})\mathbf{g}(\mathbf{X}_1)\mathbf{g}(\mathbf{x})^\tau + \mathbf{a}_2(\mathbf{X}_1, \mathbf{x})\mathbf{g}(\mathbf{x})^\tau] - E_{\mathbf{Z}_1}[\mathbf{g}(\mathbf{X}_1)a_1'(\mathbf{X}_1, \mathbf{x})^\tau + \mathbf{a}_2'(\mathbf{X}_1, \mathbf{x})], \tag{A.2}$$

the expectation is taken with respect to $\mathbf{Z}_1 = (Y_1, \mathbf{X}_1)$, and

$$a_1(\mathbf{x}_i, \mathbf{x}_j) = \int b_1(\mathbf{z}_i, \mathbf{z}_j) f_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) f_{Y|\mathbf{X}}(y_j|\mathbf{x}_j) \mathrm{d}y_i \mathrm{d}y_j, \; a_1'(\mathbf{x}_i, \mathbf{x}_j) = \frac{\partial a_1(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_j},$$
$$\mathbf{a}_2(\mathbf{x}_i, \mathbf{x}_j) = \int \mathbf{b}_2(\mathbf{z}_i, \mathbf{z}_j) f_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) f_{Y|\mathbf{X}}(y_j|\mathbf{x}_j) \mathrm{d}y_i \mathrm{d}y_j, \; \mathbf{a}_2'(\mathbf{x}_i, \mathbf{x}_j) = \frac{\partial \mathbf{a}_2(\mathbf{x}_i, \mathbf{x}_j)}{\partial \mathbf{x}_j^\tau}.$$

Theorem 6. *Assume the conditions* (A1)–(A4) *(stated in Appendix A.2) and the conditions* (A5g)–(A9g) *(given below) hold. If (a) $n \to \infty$, $h \to 0$, $b \to 0$, and $b^{-1}h \to 0$; (b) for some $\varepsilon > 0$, $b^4 n^{1-\varepsilon} h^{2p+2} \to \infty$; and (c) $nh^{2s-2} \to 0$, then* $\text{vec}(\hat{\mathcal{M}}_n)$ *asymptotically follows a multivariate normal distribution,*

$$\sqrt{n}\,(\text{vec}(\hat{\mathcal{M}}_n) - \text{vec}(\mathcal{M})) \xrightarrow{\mathcal{L}} N(0, \boldsymbol{\Sigma}), \quad \text{as } n \to \infty,$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $vec(\mathcal{R}(\mathbf{Z}) + \mathcal{R}(\mathbf{Z})^\tau)$.

Given below are the technical conditions (A5g)–(A9g) required by Theorem 6.

(A5g) $a_1(\mathbf{x}_1, \mathbf{x}_2)$ and $\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2)$ have continuous derivatives on $\Omega_0 \times \Omega_0 \subset \Omega \times \Omega$, where $\Omega \times \Omega - \Omega_0 \times \Omega_0$ is a set of measure 0.

(A6g) The following expectations are finite.

$$E\|b_1(\mathbf{Z}_1, \mathbf{Z}_2)\|^2, \quad E\|b_1(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{g}(\mathbf{X}_1)\|^2, \quad E\|b_1(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{g}(\mathbf{X}_1)\mathbf{g}(\mathbf{X}_2)^\tau\|^2,$$
$$E\|\mathcal{B}_3(\mathbf{Z}_1, \mathbf{Z}_2)\|^2, \quad E\|\mathbf{b}_2(\mathbf{Z}_1, \mathbf{Z}_2)\|^2, \quad E\|\mathbf{b}_2(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{g}(\mathbf{X}_2)^\tau\|^2.$$

(A7g) Let $\bar{I}_{12} = 1 - I_1 I_2$. As $n \to \infty$, the following statements hold.

$$E[b_1(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{g}(\mathbf{X}_1)\mathbf{g}(\mathbf{X}_2)^\tau \bar{I}_{12}] = o(n^{-1/2}),$$
$$E[\mathbf{b}_2(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{g}(\mathbf{X}_2)^\tau \bar{I}_{12}] = o(n^{-1/2}),$$
$$E[\mathcal{B}_3(\mathbf{Z}_1, \mathbf{Z}_2)\bar{I}_{12}] = o(n^{-1/2}).$$

(A8g) The functions $f_{\mathbf{X}}(\mathbf{x})$, $f_{\mathbf{X}}'(\mathbf{x})$, $a_1'(\mathbf{x}_1, \mathbf{x})$, $\mathbf{a}_2'(\mathbf{x}_1, \mathbf{x})$, $a_1(\mathbf{x}_1, \mathbf{x})\mathbf{g}(\mathbf{x})$, and $\mathbf{a}_2(\mathbf{x}_1, \mathbf{x})\mathbf{g}(\mathbf{x})^\tau$ are locally Lipschitz continuous in $\mathbf{x}$ with modulus ω_f, $\omega_{f'}$, $\omega_{a_1'}$, $\omega_{a_2'}$, $\omega_{a_1 g}$, and $\omega_{a_2 g}$, respectively. We further assume that the following expectations are finite:

(a) $E\|\mathbf{b}_2(\mathbf{Z}_1, \mathbf{Z}_2)H(\mathbf{X}_2)^\tau\|^2$ and $E\|b_1(\mathbf{Z}_1, \mathbf{Z}_2)H(\mathbf{X}_1)H^*(\mathbf{X}_2)^\tau\|^2$ for H and $H^* \in \{\mathbf{g}, f_{\mathbf{X}}^{-1}\omega_{f'}, \mathbf{g}f_{\mathbf{X}}^{-1}\omega_f\}$;

(b) $E\|H_1(\mathbf{X}_1, \mathbf{X}_2)\|^2$ for $H_1 \in \{\mathbf{g}\omega_{a_1'}, \omega_{a_2'}, \mathbf{g}\omega_{a_1 g}, \omega_{a_2 g}\}$;

(c) $E\|H_2(\mathbf{X}_1, \mathbf{X}_2)f^*(\mathbf{X}_2)\|^2$ for $H_2 \in \{\omega_{a_1'}, a_1', \omega_{a_1 g}, a_1\mathbf{g}\}$ and $f^* \in \{f_{\mathbf{X}}', \omega_{f'}, \mathbf{g}f_{\mathbf{X}}, \mathbf{g}\omega_f\}$.

(A9g) Let $A_n = \{(\mathbf{x}_1, \mathbf{x}_2) : f_{\mathbf{X}}(\mathbf{x}_1) > b, f_{\mathbf{X}}(\mathbf{x}_2) > b\}$. The following integrals are bounded as $n \to \infty$.

$$\int_{A_n} \|H_3(\mathbf{x}_1, \mathbf{x}_2) f_{\mathbf{X}}(\mathbf{x}_1) f^*(\mathbf{x}_2)\| \, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2, \qquad \int_{A_n} \|H_4(\mathbf{x}_1, \mathbf{x}_2) f^*(\mathbf{x}_1) f^{**}(\mathbf{x}_2)\| \, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2$$

for $H_3 \in \{a_1\mathbf{g}, \mathbf{a}_2, h\mathbf{a}_2\mathbf{g}^\tau, ha_1\mathbf{g}\mathbf{g}^\tau\}$, $H_4 \in \{h^{s-1}a_1, h^s a_1\mathbf{g}, h^{s+1}a_1\mathbf{g}\mathbf{g}^\tau\}$, and f^* and $f^{**} \in \{f_{\mathbf{X}}^{(s)}, h^\gamma c\}$.

A.3.1. A lemma

Lemma 5. *Suppose $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ is an iid sample. Consider a general V-statistic $\mathbf{v}_n = n^{-m}\sum_{i_1=1}^n \cdots \sum_{i_m=1}^n \mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m})$, where $\mathbf{p}_n$ is symmetric under the permutation of its m variables. Let $\hat{\mathbf{u}}_n = \boldsymbol{\theta}_n + \frac{m}{n}\sum_{i=1}^n[\mathbf{r}_n(\mathbf{z}_i) - \boldsymbol{\theta}_n]$, where $\boldsymbol{\theta}_n = E[\mathbf{p}_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)]$ and $\mathbf{r}_n(\mathbf{z}_i) = E[\mathbf{p}_n(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)|\mathbf{Z}_i = \mathbf{z}_i]$. If $E\|\mathbf{p}_n(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_m})\|^2 = o(n)$ for any $i_1, \ldots, i_m \in \{1, \ldots, m\}$, then $\sqrt{n}(\mathbf{v}_n - \hat{\mathbf{u}}_n) = o_p(1)$.*

Proof of Lemma 5. The proof consists of two steps. The first step is to show that $\sqrt{n}(\mathbf{u}_n - \hat{\mathbf{u}}_n) = o_p(1)$ where $\mathbf{u}_n$ is a U-statistic defined by $\mathbf{u}_n = \binom{n}{m}^{-1}\sum_{i_1<i_2<\cdots<i_m}\mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m})$; and the second step is to show that $\sqrt{n}(\mathbf{v}_n - \mathbf{u}_n) = o_p(1)$.

The proof of the first step is the direct generalization of Lemma 3.1 in [10], where $m = 2$. Thus the detail is omitted and we only focus on proving $\sqrt{n}(\mathbf{v}_n - \mathbf{u}_n) = o_p(1)$. In fact,

$$\mathbf{v}_n - \mathbf{u}_n = \left(m!\, n^{-m} - \binom{n}{m}^{-1}\right)\sum_{i_1<\cdots<i_m}\mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}) + n^{-m}\sum_{i_1,\ldots,i_m \text{ are not distinct}}\mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}).$$

Notice that $m!\, n^{-m} - \binom{n}{m}^{-1} = O(n^{-m-1})$ and the number of summands in the second term above is $\binom{m}{2}\binom{n}{1}\binom{n}{m-2} + \cdots = O(n^{m-1})$. Because $E\|\mathbf{p}_n(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_m})\|^2 = o(n)$, we have $\mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}) = o_p(n^{1/2})$. Thus,

$$\mathbf{v}_n - \mathbf{u}_n = O(n^{-m-1})O(n^m)o_p(n^{1/2}) + n^{-m}O(n^{m-1})o_p(n^{1/2}) = o_p(n^{-1/2}).$$

Therefore, $\sqrt{n}(\mathbf{v}_n - \mathbf{u}_n) = o_p(1)$. The lemma follows by combining the two steps proven above. □

Remark 2. When $\mathbf{p}_n$ is not symmetric, define $\mathbf{p}_n^* = (m!)^{-1}\sum \mathbf{p}_n(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m})$, where the summation is over all different permutations of $(i_1, \ldots, i_m)$. Then the lemma holds for $\mathbf{p}_n^*$.

A.3.2. Proof of Theorem 6

The key step to prove Theorem 6 is to show that $\hat{\mathcal{M}}_n$ has the expansion (A.1). Once the expansion is obtained, the asymptotic normality of $\hat{\mathcal{M}}_n$ directly follows from the central limit theorem. Similar to the proof of Theorem 3.1 in [9], we divide the proof of Theorem 6 into the following four steps.

(1) Linearization: $\sqrt{n}(\bar{\mathcal{M}}_n - \tilde{\mathcal{M}}_n) = o_p(1)$, where $\bar{\mathcal{M}}_n$ is obtained by replacing $\hat{I}_i$ in $\hat{\mathcal{M}}_n$ with $I_i = I_{[f_{\mathbf{X}}(\mathbf{x}_i)>b]}$, and $\tilde{\mathcal{M}}_n$ is obtained by replacing $\hat{\mathbf{g}}(\mathbf{x}_i)$ in $\bar{\mathcal{M}}_n$ with $\tilde{\mathbf{g}}(\mathbf{x}_i)$, where

$$\tilde{\mathbf{g}}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \frac{\hat{f}_h'(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} - \frac{\hat{f}_h(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\mathbf{g}(\mathbf{x}).$$

Notice that the true density function $f_{\mathbf{X}}(\mathbf{x})$ appears in the denominators in $\tilde{\mathbf{g}}$.
(2) Asymptotic normality: $\sqrt{n}(\tilde{\mathcal{M}}_n - E(\tilde{\mathcal{M}}_n))$ asymptotically follows normal distribution.
(3) Asymptotic bias: $\sqrt{n}(E(\tilde{\mathcal{M}}_n) - \mathcal{M}) = o(1)$.
(4) Trimming effect: $\sqrt{n}(\hat{\mathcal{M}}_n - \bar{\mathcal{M}}_n) = o_p(1)$.

The asymptotic normality of $\hat{\mathcal{M}}_n$ follows by combining the above four steps together.

A.3.2.1. Linearization. Because $nh^{2s-2} \to 0$ as $n \to \infty$, the pointwise mean squared errors of $\hat{f}_h$ and $\hat{f}_h'$ are dominated by their variances. Since the set $\{\mathbf{x} \mid f_{\mathbf{X}}(\mathbf{x}) \geq b\}$ is compact and $b^{-1}h \to 0$, for any $\varepsilon > 0$, we have [26],

$$\sup_{\mathbf{x}} |\hat{f}_h(\mathbf{x}) - f_{\mathbf{X}}(\mathbf{x})| I_{[f_{\mathbf{X}}(\mathbf{x})>b]} = O_p[(n^{1-(\varepsilon/2)}h^p)^{-1/2}],$$
$$\sup_{\mathbf{x}} \|\hat{f}_h'(\mathbf{x}) - f_{\mathbf{X}}'(\mathbf{x})\| I_{[f_{\mathbf{X}}(\mathbf{x})>b]} = O_p[(n^{1-(\varepsilon/2)}h^{p+2})^{-1/2}].$$

Let c_f be a constant such that $\sup|\hat{f}_h - f_{\mathbf{X}}| I_{[f_{\mathbf{X}}(\mathbf{x})>b]} \leq c_f(n^{1-\varepsilon/2}h^p)^{-1/2}$ holds with high probability. Denote $c_n = c_f (n^{1-\varepsilon/2}h^p)^{-1/2}$. Then $\hat{f}_h(\mathbf{x}) \geq b - c_n$ holds with high probability for $\mathbf{x}$ such that $f_{\mathbf{X}}(\mathbf{x}) > b$. Notice that $b^{-1}c_n \to 0$, because

$[b^2(n^{1-\varepsilon/2}h^p)]^2 \to \infty$. We write $\bar{\mathcal{M}}_n$ explicitly as follows.

$$\begin{aligned}\bar{\mathcal{M}}_n &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n b_1(\mathbf{z}_i,\mathbf{z}_j)\hat{\mathbf{g}}(\mathbf{x}_i)\hat{\mathbf{g}}(\mathbf{x}_j)^\tau I_iI_j + n^{-2}\sum_{i=1}^n\sum_{j=1}^n \mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)\hat{\mathbf{g}}(\mathbf{x}_j)^\tau I_iI_j\\ &\quad + n^{-2}\sum_{i=1}^n\sum_{j=1}^n \hat{\mathbf{g}}(\mathbf{x}_i)\mathbf{b}_2(\mathbf{z}_j,\mathbf{z}_i)^\tau I_iI_j + n^{-2}\sum_{i=1}^n\sum_{j=1}^n \mathcal{B}_3(\mathbf{z}_i,\mathbf{z}_j)I_iI_j\\ &= \bar{\mathcal{M}}_{n,1}+\bar{\mathcal{M}}_{n,2}+\bar{\mathcal{M}}_{n,3}+\bar{\mathcal{M}}_{n,4}.\end{aligned}$$

Let $\tilde{\mathcal{M}}_{n,i}$ be the result of replacing $\hat{\mathbf{g}}(\mathbf{x}_j)$ by $\tilde{\mathbf{g}}(\mathbf{x}_j)$, for $i = 1, 2, 3$ and 4. Because $\bar{\mathcal{M}}_{n,4}$ does not involve $\hat{\mathbf{g}}$, $\bar{\mathcal{M}}_{n,4}$ and $\tilde{\mathcal{M}}_{n,4}$ are the same. The other three terms all contain $\hat{\mathbf{g}}$. Since these terms can be treated similarly, we only include the treatment of $\bar{\mathcal{M}}_{n,2}$ below. Because

$$\hat{\mathbf{g}}(\mathbf{x}_j)-\tilde{\mathbf{g}}(\mathbf{x}_j) = \frac{[f_{\mathbf{X}}(\mathbf{x}_j)-\hat{f}_h(\mathbf{x}_j)][\hat{f}_h'(\mathbf{x}_j)-f_{\mathbf{X}}'(\mathbf{x}_j)]}{\hat{f}_h(\mathbf{x}_j)f_{\mathbf{X}}(\mathbf{x}_j)} + \frac{[f_{\mathbf{X}}(\mathbf{x}_j)-\hat{f}_h(\mathbf{x}_j)]^2}{\hat{f}_h(\mathbf{x}_j)f_{\mathbf{X}}(\mathbf{x}_j)}\mathbf{g}(\mathbf{x}_j)$$

we have

$$\begin{aligned}\|\sqrt{n}(\bar{\mathcal{M}}_{n,2}-\tilde{\mathcal{M}}_{n,2})\| &\le n^{-2}\sum_{i=1}^n\sum_{j=1}^n I_iI_j\cdot\|\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)\|\frac{\sqrt{n}}{b(b-c_n)}\{\sup|f_{\mathbf{X}}-\hat{f}_h|I\}\{\sup\|\hat{f}_h'-f_{\mathbf{X}}'\|I\}\\ &\quad + n^{-2}\sum_{i=1}^n\sum_{j=1}^n I_iI_j\cdot\|\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_j)^\tau\|\frac{\sqrt{n}}{b(b-c_n)}\{\sup|f_{\mathbf{X}}-\hat{f}_h|I_j\}^2\\ &= O_p(b^{-2}n^{-(1-\varepsilon)/2}h^{-p-1})+O_p(b^{-2}n^{-(1-\varepsilon)/2}h^{-p}) = o_p(1).\end{aligned}$$

Note that $n^{-2}\sum_{i=1}^n\sum_{j=1}^n\|\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)\|I_iI_j$ and $n^{-2}\sum_{i=1}^n\sum_{j=1}^n\|\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_j)^\tau\|I_iI_j$ are bounded in probability, because $E\|\mathbf{b}_2(\mathbf{Z}_1,\mathbf{Z}_2)\|^2$ and $E\|\mathbf{b}_2(\mathbf{Z}_1,\mathbf{Z}_2)\mathbf{g}(\mathbf{x}_2)^\tau\|^2$ are finite; $b^{-2}n^{-(1-\varepsilon)/2}h^{-p-1} = o(1)$ because $b^4n^{1-\varepsilon}h^{2p+2}\to\infty$. Similarly, we can show that both $\sqrt{n}(\bar{\mathcal{M}}_{n,1}-\tilde{\mathcal{M}}_{n,1})$ and $\sqrt{n}(\bar{\mathcal{M}}_{n,3}-\tilde{\mathcal{M}}_{n,3})$ are of $o_p(1)$. Therefore, $\sqrt{n}(\bar{\mathcal{M}}_n-\tilde{\mathcal{M}}_n) = o_p(1)$.

A.3.2.2. Asymptotic normality. We need to express $\tilde{\mathcal{M}}_n - E(\tilde{\mathcal{M}}_n)$ as the sum of the average of n iid random matrices and a negligible term of order $o_p(n^{-1/2})$ by applying Lemma 5. Then the asymptotic normality directly follows from the central limit theorem. Some calculations lead to the following expression of $\tilde{\mathcal{M}}_n$,

$$\begin{aligned}\tilde{\mathcal{M}}_n &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n\Big\{\mathcal{U}_1(\mathbf{z}_i,\mathbf{z}_j)I_iI_j + \mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j)\hat{f}_h'(\mathbf{x}_j)^\tau I_iI_j + \hat{f}_h'(\mathbf{x}_i)\mathbf{u}_2(\mathbf{z}_j,\mathbf{z}_i)^\tau I_iI_j - \mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_j)^\tau\hat{f}_h(\mathbf{x}_j)I_iI_j\\ &\quad - \mathbf{g}(\mathbf{x}_i)\mathbf{u}_2(\mathbf{z}_j,\mathbf{z}_i)^\tau\hat{f}_h(\mathbf{x}_i)I_iI_j + u_3(\mathbf{z}_i,\mathbf{z}_j)\hat{f}_h'(\mathbf{x}_i)\hat{f}_h'(\mathbf{x}_j)^\tau I_iI_j + u_3(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_j)^\tau\hat{f}_h(\mathbf{x}_i)\hat{f}_h(\mathbf{x}_j)I_iI_j\\ &\quad - \hat{f}_h'(\mathbf{x}_i)\hat{f}_h(\mathbf{x}_j)u_3(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_j)^\tau I_iI_j - u_3(\mathbf{z}_j,\mathbf{z}_i)\mathbf{g}(\mathbf{x}_i)\hat{f}_h(\mathbf{x}_i)\hat{f}_h'(\mathbf{x}_j)^\tau I_iI_j\Big\},\end{aligned}$$

where

$$\begin{aligned}\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j) &= b_1(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_i)f_{\mathbf{X}}^{-1}(\mathbf{x}_j)+\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)f_{\mathbf{X}}^{-1}(\mathbf{x}_j),\\ u_3(\mathbf{z}_i,\mathbf{z}_j) &= b_1(\mathbf{z}_i,\mathbf{z}_j)f_{\mathbf{X}}^{-1}(\mathbf{x}_i)f_{\mathbf{X}}^{-1}(\mathbf{x}_j).\end{aligned}$$

In what follows, we only discuss the first two terms in the expression of $\tilde{\mathcal{M}}_n$ above, and the remaining terms can be treated similarly.

Applying Lemma 5 to the first term of $\tilde{\mathcal{M}}_n$,

$$\begin{aligned}&n^{-2}\sum_{i=1}^n\sum_{j=1}^n\mathcal{U}_1(\mathbf{z}_i,\mathbf{z}_j)I_iI_j - E[\mathcal{U}_1(\mathbf{Z}_1,\mathbf{Z}_2)I_1I_2]\\ &\quad = n^{-1}\sum_{i=1}^n\{E[\mathcal{U}_1(\mathbf{z}_i,\mathbf{Z}_1)I_iI_1]+E[\mathcal{U}_1(\mathbf{Z}_1,\mathbf{z}_i)I_1I_i]-2E[\mathcal{U}_1(\mathbf{Z}_1,\mathbf{Z}_2)I_1I_2]\}+o_p(n^{-1/2})\\ &\quad = n^{-1}\sum_{i=1}^n\{E[\mathcal{U}_1(\mathbf{z}_i,\mathbf{Z}_1)]+E[\mathcal{U}_1(\mathbf{Z}_1,\mathbf{z}_i)]-2E[\mathcal{U}_1(\mathbf{Z}_1,\mathbf{Z}_2)]\}+o_p(n^{-1/2}).\end{aligned}$$

For the second term of $\tilde{\mathcal{M}}_n$,

$$n^{-2}\sum_{i=1}^n\sum_{j=1}^n\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j)\hat{f}_h'(\mathbf{x}_j)^\tau I_iI_j = n^{-3}h^{-p-1}\sum_{i=1}^n\sum_{j=1}^n\sum_{\ell=1}^n\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j)K'\left(\frac{\mathbf{x}_j-\mathbf{x}_\ell}{h}\right)^\tau I_iI_j,$$

which is a general V-statistic. We need to verify $E\|h^{-p-1}\mathbf{u}_2(\mathbf{Z}_i,\mathbf{Z}_j)K'(\frac{\mathbf{X}_j-\mathbf{X}_\ell}{h})^\tau I_iI_j\|^2 = o(n)$ before applying Lemma 5.

$$\begin{aligned}E\|h^{-p-1}\mathbf{u}_2(\mathbf{Z}_i,\mathbf{Z}_j)K'\left(\frac{\mathbf{X}_j-\mathbf{X}_\ell}{h}\right)^\tau I_iI_j\|^2 &\le b^{-2}h^{-p-2}\int\|(\mathbf{b}_2(\mathbf{z}_i,\mathbf{z}_j)+b_1(\mathbf{z}_i,\mathbf{z}_j)\mathbf{g}(\mathbf{x}_i))K'(\mathbf{u})^\tau\|^2\\ &\quad\times f_{Y,\mathbf{X}}(y_i,\mathbf{x}_i)f_{Y,\mathbf{X}}(y_j,\mathbf{x}_j)f_{\mathbf{X}}(\mathbf{x}_j-h\mathbf{u})\mathrm{d}\mathbf{x}_i\mathrm{d}\mathbf{x}_j\mathrm{d}\mathbf{u}\mathrm{d}y_i\mathrm{d}y_j\\ &= O(b^{-2}h^{-p-2}) = O(n(b^2nh^{p+2})^{-1}) = o(n).\end{aligned}$$

Notice that K is symmetric, so $K'(0)=0$. Then $E\|h^{-p-1}\mathbf{u}_2(\mathbf{Z}_i,\mathbf{Z}_j)K'(0)^\tau I_iI_j\| = 0$, which can be considered of order $o(n)$. Applying Lemma 5, we can write

$$\begin{aligned}&n^{-2}\sum_{i=1}^n\sum_{j=1}^n\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_j)\hat{f}_h'(\mathbf{x}_j)^\tau I_iI_j - h^{-p-1}E\left[\mathbf{u}_2(\mathbf{Z}_1,\mathbf{Z}_2)K'\left(\frac{\mathbf{X}_2-\mathbf{X}_3}{h}\right)^\tau I_1I_2\right]\\ &\quad= n^{-1}\sum_{i=1}^n\{E[\mathcal{W}_2(\mathbf{z}_i,\mathbf{Z}_1,\mathbf{Z}_2)]-E[\mathcal{W}_2(\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3)]\}+o_p(n^{-1/2}),\end{aligned}\tag{A.3}$$

where

$$h^{p+1}\,\mathcal{W}_2(\mathbf{z}_i,\mathbf{z}_1,\mathbf{z}_2) = \mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_1)K'\left(\frac{\mathbf{x}_1-\mathbf{x}_2}{h}\right)^\tau I_iI_1+\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_i)K'\left(\frac{\mathbf{x}_i-\mathbf{x}_2}{h}\right)^\tau I_1I_i+\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_2)K'\left(\frac{\mathbf{x}_2-\mathbf{x}_i}{h}\right)^\tau I_1I_2.$$

Next, we want to show that $E[\mathcal{W}_2(\mathbf{z}_i,\mathbf{Z}_1,\mathbf{Z}_2)]$ can be approximated by an expression without I_i's, which therefore does not depend on n. Notice that $f_{\mathbf{X}}'$ satisfies local Lipschitz condition (A8g) and by applying Cauchy–Schwarz inequality $E\|\mathbf{u}_2(\mathbf{Z}_1,\mathbf{Z}_2)\omega_{f'}\|^2$ is finite. For the first term of $E[\mathcal{W}_2(\mathbf{z}_i,\mathbf{Z}_1,\mathbf{Z}_2)]$, we have

$$\begin{aligned}&\int h^{-p-1}\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_1)K'\left(\frac{\mathbf{x}_1-\mathbf{x}_2}{h}\right)^\tau I_iI_1f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{\mathbf{X}}(\mathbf{x}_2)\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\mathrm{d}\mathbf{x}_2\\ &\quad=\int\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_1)K(\mathbf{u})I_iI_1f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{\mathbf{X}}'(\mathbf{x}_1-h\mathbf{u})^\tau\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\mathrm{d}\mathbf{u}\\ &\quad=\int\mathbf{u}_2(\mathbf{z}_i,\mathbf{z}_1)f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{\mathbf{X}}'(\mathbf{x}_1)^\tau\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1+o_p(1).\end{aligned}\tag{A.4}$$

Similarly for the second term of $E[\mathcal{W}_2(\mathbf{z}_i,\mathbf{Z}_1,\mathbf{Z}_2)]$,

$$\begin{aligned}&\int h^{-p-1}\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_i)K'\left(\frac{\mathbf{x}_i-\mathbf{x}_2}{h}\right)^\tau I_iI_1f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{\mathbf{X}}(\mathbf{x}_2)\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\mathrm{d}\mathbf{x}_2\\ &\quad=\int\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_i)f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\cdot f_{\mathbf{X}}'(\mathbf{x}_i)^\tau+o_p(1).\end{aligned}\tag{A.5}$$

Now we discuss the third term of $E[\mathcal{W}_2(\mathbf{z}_i,\mathbf{Z}_1,\mathbf{Z}_2)]$. Let $b^*=\sup_{\mathbf{x},\mathbf{u}}\{f_{\mathbf{X}}(\mathbf{x}+h\mathbf{u})\mid f_{\mathbf{X}}(\mathbf{x})=b,|\mathbf{u}|\le 1\}$ and $I_i^*=I_{[f_{\mathbf{X}}(\mathbf{x}_i)>b^*]}$. When $I_i^*=1$, it is always true that $I_{[f_{\mathbf{X}}(\mathbf{x}_i+h\mathbf{u})>b]}=1$.

$$\begin{aligned}&\int h^{-p-1}\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_2)K'\left(\frac{\mathbf{x}_2-\mathbf{x}_i}{h}\right)^\tau I_1I_2f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{Y,\mathbf{X}}(y_2,\mathbf{x}_2)\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\mathrm{d}\mathbf{x}_2\mathrm{d}y_2\\ &\quad=-I^*\int[\mathbf{a}_2'(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})+\mathbf{g}(\mathbf{x}_1)a_1'(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})]I_1K(\mathbf{u})f_{\mathbf{X}}(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{u}+(1-I^*)\mathcal{A}(\mathbf{x}_i;h,b),\end{aligned}$$

where

$$\mathcal{A}(\mathbf{x}_i;h,b)=\int h^{-1}[\mathbf{a}_2(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})+a_1(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})\mathbf{g}(\mathbf{x}_1)]K'(\mathbf{u})^\tau I_1I_{[f_{\mathbf{X}}(\mathbf{x}_i+h\mathbf{u})>b]}f_{\mathbf{X}}(\mathbf{x}_1)\,\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{u}.$$

Then

$$\begin{aligned}&\int h^{-p-1}\mathbf{u}_2(\mathbf{z}_1,\mathbf{z}_2)K'\left(\frac{\mathbf{x}_2-\mathbf{x}_i}{h}\right)^\tau I_1I_2f_{Y,\mathbf{X}}(y_1,\mathbf{x}_1)f_{Y,\mathbf{X}}(y_2,\mathbf{x}_2)\,\mathrm{d}\mathbf{x}_1\mathrm{d}y_1\mathrm{d}\mathbf{x}_2\mathrm{d}y_2\\ &\qquad+\int[\mathbf{a}_2'(\mathbf{x}_1,\mathbf{x}_i)+\mathbf{g}(\mathbf{x}_1)a_1'(\mathbf{x}_1,\mathbf{x}_i)^\tau]f_{\mathbf{X}}(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1\\ &\quad=-I^*\int[\mathbf{a}_2'(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})-\mathbf{a}_2'(\mathbf{x}_1,\mathbf{x}_i)]K(\mathbf{u})f_{\mathbf{X}}(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{u}\\ &\qquad-I^*\int\mathbf{g}(\mathbf{x}_1)[a_1'(\mathbf{x}_1,\mathbf{x}_i+h\mathbf{u})-a_1'(\mathbf{x}_1,\mathbf{x}_i)]^\tau K(\mathbf{u})f_{\mathbf{X}}(\mathbf{x}_1)\mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{u}\end{aligned}$$

$$+I^* \int [\mathbf{a}_2'(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u}) + \mathbf{g}(\mathbf{x}_1) a_1'(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u})^\tau] K(\mathbf{u})(1 - I_1) f_{\mathbf{X}}(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{u}$$

$$+(1 - I^*) \int [\mathbf{a}_2'(\mathbf{x}_1, \mathbf{x}_i) + \mathbf{g}(\mathbf{x}_1) a_1'(\mathbf{x}_i, \mathbf{x}_1)^\tau] f_{\mathbf{X}}(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 + (1 - I^*) \mathcal{A}(\mathbf{x}_i; h, b). \tag{A.6}$$

The first and second terms on the right-hand side of (A.6) are of order $o_p(1)$ due to assumption (A8g). The third and forth terms on the right-hand side of (A.6) are of order $o_p(1)$ because $b \to 0$. To verify that the fifth term $(1 - I^*)\mathcal{A}(\mathbf{x}_i; h, b)$ vanishes, we only need to show that the second moment of $\mathcal{A}(\mathbf{x}_i; h, b)$ is finite. Suppose the kth column of $\mathcal{A}(\mathbf{x}_i; h, b)$ is $\mathbf{a}_k(\mathbf{x}_i; h, b)$. Then applying integration by parts to $\mathbf{a}_k(\mathbf{x}_i; h, b)$, we have

$$\begin{aligned}\mathbf{a}_k(\mathbf{x}_i; h, b) = &- \int \frac{\partial}{\partial u_k} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u}) + \mathbf{g}(\mathbf{x}_1) a_1(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u})] K(\mathbf{u}) I_1 I_{[f_{\mathbf{X}}(\mathbf{x}_i + h\mathbf{u}) > b]} \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{u} \\ &- \int_D [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u}) + a_1(\mathbf{x}_1, \mathbf{x}_i + h\mathbf{u}) \mathbf{g}(\mathbf{x}_1)] K(\mathbf{u}) I_1 \, \mathrm{d}\mathbf{x}_i \mathrm{d}\mathbf{u}_{(k)},\end{aligned}$$

where D contains all $\mathbf{u}$ such that $f_{\mathbf{X}}(\mathbf{x}_i + h\mathbf{u}) = b$ and $d\mathbf{u}_{(k)} = du_1 \cdots du_{k-1} du_{k+1} \cdots du_p$. Thus the second moment of $\mathbf{a}_k(\mathbf{X}; h, b)$ exists, which further implies that the second moment of $\mathcal{A}(\mathbf{X}; h, b)$ exists. Therefore, all terms on the right-hand side of (A.6) are of order $o_p(1)$ and we obtain

$$\begin{aligned}&\int h^{-p-1} \mathbf{u}_2(\mathbf{z}_1, \mathbf{z}_2) K' \left(\frac{\mathbf{x}_2 - \mathbf{x}_i}{h}\right)^\tau I_1 I_2 f_{Y,\mathbf{X}}(y_1, \mathbf{x}_1) f_{Y,\mathbf{X}}(y_2, \mathbf{x}_2) \, \mathrm{d}\mathbf{x}_1 \mathrm{d}y_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}y_2 \\ &= - \int [\mathbf{a}_2'(\mathbf{x}_1, \mathbf{x}_i) + \mathbf{g}(\mathbf{x}_1) a_1'(\mathbf{x}_1, \mathbf{x}_i)^\tau] f_{\mathbf{X}}(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1 + o_p(1).\end{aligned} \tag{A.7}$$

Combining (A.3)–(A.5) and (A.7) together, we have

$$\begin{aligned}&n^{-2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{u}_2(\mathbf{z}_i, \mathbf{z}_j) \hat{f}_h'(\mathbf{x}_j)^\tau I_i I_j - h^{-p-1} E\left[\mathbf{u}_2(\mathbf{Z}_1, \mathbf{Z}_2) K' \left(\frac{\mathbf{X}_2 - \mathbf{X}_3}{h}\right)^\tau I_1 I_2\right] \\ &= n^{-1} \sum_{i=1}^n \{\tilde{\mathcal{W}}_2(\mathbf{z}_i) - E[\tilde{\mathcal{W}}_2(\mathbf{Z}_1)]\} + o_p(n^{-1/2}),\end{aligned}$$

where

$$\begin{aligned}\tilde{\mathcal{W}}_2(\mathbf{z}_i) = &\int \mathbf{u}_2(\mathbf{z}_i, \mathbf{z}_1) f_{Y,\mathbf{X}}(y_1, \mathbf{x}_1) f_{\mathbf{X}}'(\mathbf{x}_1)^\tau \, \mathrm{d}\mathbf{x}_1 \mathrm{d}y_1 + \int \mathbf{u}_2(\mathbf{z}_1, \mathbf{z}_i) f_{Y,\mathbf{X}}(y_1, \mathbf{x}_1) \, \mathrm{d}\mathbf{x}_1 \mathrm{d}y_1 \cdot f_{\mathbf{X}}'(\mathbf{x}_i)^\tau \\ &- \int [\mathbf{a}_2'(\mathbf{x}_1, \mathbf{x}_i) + \mathbf{g}(\mathbf{x}_1) a_2'(\mathbf{x}_1, \mathbf{x}_i)^\tau] f_{\mathbf{X}}(\mathbf{x}_1) \mathrm{d}\mathbf{x}_1.\end{aligned}$$

Following the similar arguments, we can have expansions for the other terms of $\tilde{\mathcal{M}}_n - E(\tilde{\mathcal{M}}_n)$. Collecting all these terms together, we obtain an expansion for $\tilde{\mathcal{M}}_n$,

$$\tilde{\mathcal{M}}_n - E(\tilde{\mathcal{M}}_n) = n^{-1} \sum_{i=1}^n \{\mathcal{R}(\mathbf{z}_i) + \mathcal{R}(\mathbf{z}_i)^\tau - E[\mathcal{R}(\mathbf{Z}_1) + \mathcal{R}(\mathbf{Z}_1)^\tau]\} + o_p(n^{-1/2}),$$

where $\mathcal{R}(\mathbf{z})$ is defined in (A.2).

A.3.2.3. Asymptotic bias. Notice that $\mathcal{M} = E[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{Z}_2)]$. The bias is

$$\begin{aligned}E(\tilde{\mathcal{M}}_n) - \mathcal{M} = &\{E[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{Z}_2) I_1 I_2] - \mathcal{M}\} + h^{-p-1} E[\mathbf{u}_2(\mathbf{Z}_1, \mathbf{Z}_2) {K'}_{23}^{\tau} I_1 I_2] + h^{-p-1} E[K_{13}' \mathbf{u}_2(\mathbf{Z}_2, \mathbf{Z}_1)^\tau I_1 I_2] \\ &- h^{-p} E[\mathbf{u}_2(\mathbf{Z}_1, \mathbf{Z}_2) \mathbf{g}(\mathbf{X}_2)^\tau K_{23} I_1 I_2] - h^{-p} E[\mathbf{g}(\mathbf{X}_1) \mathbf{u}_2(\mathbf{Z}_2, \mathbf{Z}_1)^\tau K_{13} I_1 I_2] \\ &+ h^{-2p-2} E[u_3(\mathbf{Z}_1, \mathbf{Z}_2) K_{13}' {K'}_{24}^{\tau} I_1 I_2] + h^{-2p} E[u_3(\mathbf{Z}_1, \mathbf{Z}_2) \mathbf{g}(\mathbf{X}_1) \mathbf{g}(\mathbf{X}_2)^\tau K_{13} K_{24} I_1 I_2] \\ &- h^{-2p-1} E[K_{13}' K_{24} u_3(\mathbf{Z}_1, \mathbf{Z}_2) \mathbf{g}(\mathbf{X}_2)^\tau I_1 I_2] - h^{-2p-1} E[u_3(\mathbf{Z}_2, \mathbf{Z}_1) \mathbf{g}(\mathbf{X}_1) K_{13} {K'}_{24}^{\tau} I_1 I_2],\end{aligned}$$

where $K_{ij} = K((\mathbf{X}_i - \mathbf{X}_j)/h)$ and $K_{ij}' = K'((\mathbf{X}_i - \mathbf{X}_j)/h)$.

For the first term of $E(\tilde{\mathcal{M}}_n) - \mathcal{M}$, $E[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{Z}_2) I_1 I_2] - \mathcal{M} = o(n^{-1/2})$ due to condition (A7g). Let ℓ denote an index set $(\ell_1, \ldots, \ell_k)$ with $\sum \ell_i = s$. For $\mathbf{u} = (u_1, \ldots, u_k)$, define $\mathbf{u}^\ell = u_1^{\ell_1} \cdots u_k^{\ell_k}$ and $f_\ell^{(s)} = \partial^\ell f_{\mathbf{X}} / (\partial \mathbf{u})^\ell$. For the second term of $E(\tilde{\mathcal{M}}_n) - \mathcal{M}$,

$$\begin{aligned}&h^{-p-1} E\left[\mathbf{u}_2(\mathbf{Z}_1, \mathbf{Z}_2) K' \left(\frac{\mathbf{X}_2 - \mathbf{X}_3}{h}\right)^\tau I_1 I_2\right] = \int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2) \mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) f_{\mathbf{X}}'(\mathbf{x}_2)^\tau \, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \\ &+ h^{s-1} \int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2) \mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) K(\mathbf{u}) \sum_\ell f_\ell^{(s)}(\xi) u^\ell \, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \mathrm{d}\mathbf{u},\end{aligned} \tag{A.8}$$

where ξ lies on the line segment between $\mathbf{x}_2$ and $\mathbf{x}_2 - h\mathbf{u}$. Consider the integral in the second term in (A.8),

$$\begin{aligned}
&\int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2)\mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) \sum f_\ell^{(s)}(\xi)\, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \\
&\quad = \sum \int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2)\mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) f_\ell^{(s)}(\mathbf{x}_2)\, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \\
&\qquad + \sum \int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2)\mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) [f_\ell^{(s)}(\xi) - f_\ell^{(s)}(\mathbf{x}_2)]\, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 = O_p(1).
\end{aligned}$$

The last equation holds because $f_{\mathbf{X}}^{(s)}$ is locally Hölder continuous and the integrals in the second term in the above expression are bounded due to condition (A9g). Thus (A.8) becomes

$$h^{-p-1} E\left[\mathbf{u}_2(\mathbf{Z}_1, \mathbf{Z}_2) K'\left(\frac{\mathbf{X}_2 - \mathbf{X}_3}{h}\right)^\tau I_1 I_2\right] = \int_{A_n} [\mathbf{a}_2(\mathbf{x}_1, \mathbf{x}_2) + a_1(\mathbf{x}_1, \mathbf{x}_2)\mathbf{g}(\mathbf{x}_1)] f_{\mathbf{X}}(\mathbf{x}_1) f'_{\mathbf{X}}(\mathbf{x}_2)^\tau\, \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 + O_p(h^{s-1}).$$

Similarly we can obtain expansions for the other terms of $E(\tilde{\mathcal{M}}_n) - \mathcal{M}$. After collecting these terms together, we find that the lower-order terms sum to zero and the bias $E(\tilde{\mathcal{M}}_n) - \mathcal{M}$ is of order $O_p(h^{s-1}) = o_p(n^{-1/2})$.

A.3.2.4. Trimming effect. Because $b^{-1}c_n \to 0$, the previous result still holds if we change $I_{[f_{\mathbf{X}}(\mathbf{x})>b]}$ to $I_{[f_{\mathbf{X}}(\mathbf{x})>b+c_n]}$. Denote $\tilde{I} = I_{[\hat{f}_h(\mathbf{x})>b]} - I_{[f_{\mathbf{X}}(\mathbf{x})>b+c_n]}$, which is equal to $I_{[f_{\mathbf{X}}(\mathbf{x})\le b+c_n; \hat{f}_h(\mathbf{x})>b]}$. (The equality holds with large probability because sup $|f_{\mathbf{X}}(\mathbf{x}) - \hat{f}_h(\mathbf{x})|I \le c_n$ with large probability.) Define

$$\tilde{I}_{ij} = I_{[\hat{f}_h(\mathbf{x}_i)>b]} I_{[\hat{f}_h(\mathbf{x}_j)>b]} - I_{[f_{\mathbf{X}}(\mathbf{x}_i)>b+c_n]} I_{[f_{\mathbf{X}}(\mathbf{x}_j)>b+c_n]} = \tilde{I}_i \tilde{I}_j + \tilde{I}_i I_{[f_{\mathbf{X}}(\mathbf{x}_j)>b+c_n]} + I_{[f_{\mathbf{X}}(\mathbf{x}_i)>b+c_n]} \tilde{I}_j,$$

so

$$\begin{aligned}
\sqrt{n}(\hat{\mathcal{M}}_n - \bar{\mathcal{M}}_n) &= n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{U}_1(\mathbf{z}_i, \mathbf{z}_j) \tilde{I}_{ij} + n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{b}_2(\mathbf{z}_i, \mathbf{z}_j)(\hat{\mathbf{g}}(\mathbf{x}_j) - \mathbf{g}(\mathbf{x}_j))^\tau \tilde{I}_{ij} \\
&\quad + n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n (\hat{\mathbf{g}}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_i)) \mathbf{b}_2(\mathbf{z}_j, \mathbf{z}_i)^\tau \tilde{I}_{ij} \\
&\quad + n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n b_1(\mathbf{z}_i, \mathbf{z}_j)(\hat{\mathbf{g}}(\mathbf{x}_i)\hat{\mathbf{g}}(\mathbf{x}_j)^\tau - \mathbf{g}(\mathbf{x}_i)\mathbf{g}(\mathbf{x}_j)^\tau) \tilde{I}_{ij}.
\end{aligned}$$

We only discuss the first terms of the foregoing expression in detail. The treatment of the remaining terms is similar to that in the linearization step. Consider

$$\begin{aligned}
E\left\| n^{-3/2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{U}_1(\mathbf{z}_i, \mathbf{z}_j) \tilde{I}_{ij} \right\|^2 &\le n^{-3} E\left(\sum_{i=1}^n \sum_{j=1}^n \|\mathcal{U}_1(\mathbf{Z}_i, \mathbf{Z}_j)\| \tilde{I}_{ij} \right)^2 \\
&= n[O_p(E[\mathcal{U}_1(\mathbf{Z}_i, \mathbf{Z}_j)\bar{I}_{ij}])]^2 = o_p(1).
\end{aligned}$$

The above equation holds because $E[\mathcal{U}_1(\mathbf{Z}_1, \mathbf{Z}_2)\bar{I}_{12}] = o(n^{-1/2})$. Thus the first term of $\sqrt{n}(\hat{\mathcal{M}}_n - \bar{\mathcal{M}}_n)$ is of order $o_p(1)$. Therefore, the effect of trimming is negligible.

A.4. Proofs of Theorems 2–5

Theorem 2 can in fact be considered a special case of Theorem 6, where

$$\begin{aligned}
b_1(\mathbf{z}_1, \mathbf{z}_2) &= \int H(y_1, v) H(y_2, v) \mathrm{d}v \int W(\mathbf{x}_1, \mathbf{u}) W(\mathbf{x}_2, \mathbf{u}) \mathrm{d}\mathbf{u}, \\
\mathbf{b}_2(\mathbf{z}_1, \mathbf{z}_2) &= \int H(y_1, v) H(y_2, v) \mathrm{d}v \int \frac{\partial W(\mathbf{x}_1, \mathbf{u})}{\partial \mathbf{x}} W(\mathbf{x}_2, \mathbf{u}_2) \mathrm{d}\mathbf{u}, \\
\mathcal{B}_3(\mathbf{z}_1, \mathbf{z}_2) &= \int H(y_1, v) H(y_2, v) \mathrm{d}v \int \frac{\partial W(\mathbf{x}_1, \mathbf{u})}{\partial \mathbf{x}} \frac{\partial W(\mathbf{x}_2, \mathbf{u})}{\partial \mathbf{x}^\tau} \mathrm{d}\mathbf{u}.
\end{aligned}$$

Applying Theorem 6 with $\mathcal{U}_1 = \mathcal{U}_{\text{ITC}}$, it can be verified that $\mathcal{R}(\mathbf{z})$ in Theorem 6 becomes $\mathcal{R}(\mathbf{z})$ in Theorem 2. Therefore, it is sufficient to show that the conditions required by Theorem 2 implies the conditions required by Theorem 6. In other words, the conditions (A5c)–(A9c) implies the conditions (A5g)–(A9g). This can be proved by applying the Cauchy–Schwarz inequality repeatedly and noticing that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent. Due to space limitation, we omit the details.

Remark 3. The only difference between Theorems 2 and 3 is that $\int H(y_1, v)H(y_2, v)\mathrm{d}v$ in Theorem 2 is replaced by Y_1Y_2 in Theorem 3. The proof of Theorem 3 is similar to that of Theorem 2, and is omitted due to limited space.

Remark 4. Both $\mathcal{M}_{\mathrm{ITCe}}$ and $\mathcal{M}_{\mathrm{ITMe}}$ can be expressed as $\mathcal{M} = E[\mathcal{U}_e(\mathbf{Z}_1, \mathbf{Z}_2)]$, where

$$\mathcal{U}_e(\mathbf{z}_1, \mathbf{z}_2) = \mathcal{B}_1(\mathbf{z}_1, \mathbf{z}_2)g_R(r_1)g_R(r_2) + \mathcal{B}_2(\mathbf{z}_1, \mathbf{z}_2)g_R(r_2) + g_R(r_1)\mathcal{B}_2(\mathbf{z}_1, \mathbf{z}_2) + \mathcal{B}_3(\mathbf{z}_1, \mathbf{z}_2),$$

and $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$ are matrix-valued functions. Given an iid sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$, an estimate of $\mathcal{M}$ is $\hat{\mathcal{M}}_n = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\mathcal{U}}_e(\mathbf{z}_i, \mathbf{z}_j)\tilde{I}_i\tilde{I}_j$, where $\hat{\mathcal{U}}_e$ is obtained by replacing g_R in $\mathcal{U}_e$ with $\tilde{f}'_{R,h}/\tilde{f}_{R,h}$. Following the similar steps as in Theorem 6, we can show that $\sqrt{n}(vec(\hat{\mathcal{M}}_n) - vec(\mathcal{M}))$ asymptotically follows a multivariate normal distribution. The technical conditions of Theorem 6 need to be modified accordingly. In fact, (A1)–(A4) become conditions imposed on the density function f_R, which is a univariate function; the conditions (A5g)–(A9g) become conditions imposed on functions involving $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$. Theorems 4 and 5 can be proved after specifying $\mathcal{B}_1$, $\mathcal{B}_2$, and $\mathcal{B}_3$ properly. Due to limited space, we do not state these technical conditions and proofs in the current article. They are available from the authors upon request.

References

[1] K.-C. Li, Sliced inverse regression for dimension reduction (with discussion), Journal of the American Statistical Association 86 (1991) 316–342.

[2] R.D. Cook, Graphics for regressions with a binary response, Journal of the American Statistical Association 91 (1996) 983–992.

[3] R.D. Cook, B. Li, Dimension reduction for conditional mean in regression, The Annals of Statistics 30 (2002) 455–474.

[4] R.D. Cook, S. Weisberg, Comments on "Sliced inverse regression for dimension reduction", Journal of the American Statistical Association 86 (1991) 328–332.

[5] B. Li, H. Zha, F. Chiaromonte, Contour regression: A general approach to dimension reduction, The Annals of Statistics 33 (2005) 1580–1616.

[6] M. Hristache, A. Juditsky, J. Polzehl, V. Spokoiny, Structure adaptive approach for dimension reduction, The Annals of Statistics 29 (2001) 1537–1566.

[7] Y. Xia, H. Tong, W.K. Li, L.-X. Zhu, An adaptive estimation of dimension reduction space (with discussion), Journal of the Royal Statistical Society, Series B: Statistical Methodology 64 (2002) 363–410.

[8] K.-C. Li, On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma, Journal of the American Statistical Association 87 (1992) 1025–1039.

[9] W. Härdle, T.M. Stoker, Investigating smooth multiple regression by the method of average derivatives, Journal of the American Statistical Association 84 (1989) 986–995.

[10] J.L. Powell, J.H. Stock, T.M. Stoker, Semiparametric estimation of index coefficients, Econometrica 57 (1989) 1403–1430.

[11] W. Härdle, J. Hart, J.S. Marron, A.B. Tsybakov, Bandwidth choice for average derivative estimation, Journal of the American Statistical Association 87 (1992) 218–226.

[12] W. Härdle, P. Hall, H. Ichimura, Optimal smoothing in single-index models, The Annals of Statistics 21 (1993) 157–178.

[13] Y. Zhu, P. Zeng, Fourier methods for estimating the central subspace and the central mean subspace in regression, Journal of the American Statistical Association 101 (2006) 1638–1651.

[14] A.M. Samarov, Exploring regression structure using nonparametric functional estimation, Journal of the American Statistical Association 88 (1993) 836–847.

[15] B.W. Silverman, Density Estimation For Statistics And Data Analysis, Chapman & Hall Ltd, 1986.

[16] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman & Hall Ltd, 1995.

[17] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, J.A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, Johns Hopkins University Press, 1993.

[18] K.T. Fang, Y.T. Zhang, Generalized Multivariate Analysis, 1990, Springer-Verlag, Berlin and Science Press, Beijing.

[19] R.D. Cook, C.J. Nachtsheim, Reweighting to achieve elliptically contoured covariates in regression, Journal of the American Statistical Association 89 (1994) 592–599.

[20] P. Hall, K.-C. Li, On almost linearity of low dimensional projections from high dimensional data, The Annals of Statistics 21 (1993) 867–889.

[21] Ye Z, R.E. Weiss, Using the bootstrap to select one of a new class of dimension reduction methods, Journal of the American Statistical Association 98 (2003) 968–979.

[22] R.D. Cook, X. Yin, Dimension reduction and visualization in discriminant analysis (with discussion), Australian and New Zealand Journal of Statistics 43 (2001) 147–199.

[23] D. Basu, C.A.B. Pereira, Conditional independence in statistics, Sankhyā, Series A 45 (1983) 324–337.

[24] P. Billingsley, Probability and Measure, John Wiley and Sons, 1995.

[25] G.B. Folland, Fourier Analysis and Its Applications, Brooks/Cole, 1992.

[26] G. Collomb, W. Härdle, Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations, Stochastic Processes and Their Applications 23 (1986) 77–89.