Successive direction extraction for estimating the central subspace in a multiple-index regression

Xiangrong Yin* Bing Li† R. Dennis Cook‡

September 10, 2007

Abstract

In this paper we propose a dimension reduction method to estimate the directions in a multiple-index regression based on information extraction. This extends the recent work of Yin and Cook (2005) who introduced the method and used it to estimate the direction in a single-index regression. While a formal extension seems conceptually straightforward, there is a fundamentally new aspect of our extension: We are able to show that, under the assumption of elliptical predictors, the estimation of multiple index regressions can be decomposed into successive single-index estimation problems. This significantly reduces the computational complexity, because the nonparametric procedure involves only a one-dimensional search at each stage. In addition, we developed a permutation test to assist in estimating the dimension of a multiple-index regression.

Key Words: Dimension-reduction subspaces; Permutation test; Regression graphics; Sufficient dimension reduction.

1 Introduction

In simple regression a 2D plot of the response Y versus the predictor $\mathbf{X}$ displays all the sample information, and can be quite helpful for gaining insights about the data and for guiding the choice of a first model. Such straightforward graphical displays of all

*Department of Statistics, 204 Statistics Building, University of Georgia, Athens, GA 30602.

†Department of Statistics, Penn State University, 326 Thomas Building, University Park, PA 16802.
This work is supported in part by National Science Foundation grants DMS-0204662 and DMS-0405681. He would like to thank Anand Vidyashankar's support to visit UGA where part of this work was done.

‡School of Statistics, 224 Church Street, S.E., 313 Ford Hall, University of Minnesota, St. Paul, MN 55455. This work was supported in part by National Science Foundation grants DMS-0405360 and 0704098.

the data are not generally possible with many predictors, but informative displays are still possible in situations where we can find low-dimensional views, the only ones that are possible in practice, that provide "sufficient" information about the regression. In regression graphics we seek to facilitate visualization of the data by reducing the dimension of the $p \times 1$ predictor vector $\mathbf{X}$ without loss of information on the regression and without requiring a pre-specified parametric model. We called this sufficient dimension reduction, borrowing terminology from classical statistics. Sufficient dimension reduction leads naturally to the idea of a *sufficient summary plot* that contains all of the information on the regression that is available from the sample.

In Section 2, we extend the information extraction method suggested by Yin and Cook (2005) to multiple index regressions. In Section 3, we demonstrate that for a regression with elliptical predictors the multiple-index optimization can be decomposed into successive single-index optimizations without loss of information. A permutation test is suggested in Section 4 to assist in estimating the minimal number of indices. In Section 5, we establish the consistency of the proposed estimator. Finally we present examples in Section 6 and further discussion in Section 7. Proofs are in the Appendix. In the rest of this section we review the concept of sufficient dimension reduction, and set the context for the subsequent results.

We assume throughout that the scalar response Y and the $p \times 1$ vector of predictors $\mathbf{X}$ are defined on a common probability space, and that the data $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are iid observations on $(Y, \mathbf{X})$ with $\mathrm{Var}(\mathbf{X}) > 0$. The notation $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{Z}$ means that the random vectors $\mathbf{U}$ and $\mathbf{V}$ are independent given any value for the random vector $\mathbf{Z}$. Subspaces will be denoted by $\mathcal{S}$ and, for a $t \times u$ matrix $\mathbf{B}$, $\mathcal{S}(\mathbf{B})$ means the subspace of $\mathbb{R}^t$ spanned by the columns of $\mathbf{B}$. $P_{\mathbf{B}}$ denotes the projection operator for $\mathcal{S}(\mathbf{B})$ with respect to the usual inner product and $Q_{\mathbf{B}} = I - P_{\mathbf{B}}$.

1.1 Sufficient dimension reduction

Let $\mathbf{B}$ denote a fixed $p \times q$, $q \leq p$, matrix so that

$$Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}. \tag{1}$$

Statement (1) holds when $\mathbf{B}$ is replaced with any matrix whose columns form a basis for $\mathcal{S}(\mathbf{B})$. Thus, (1) is appropriately viewed as a statement about $\mathcal{S}(\mathbf{B})$, which is called a *dimension-reduction subspace* (DRS) for the regression of Y on $\mathbf{X}$. Let $\mathcal{S}_{Y|\mathbf{X}}$ denote the intersection of all DRS's. While $\mathcal{S}_{Y|\mathbf{X}}$ is always a subspace, it is not necessarily a DRS. However, under various reasonable conditions $\mathcal{S}_{Y|\mathbf{X}}$ is a DRS (Cook 1994, 1996, 1998), and it is then called the the *central subspace* (CS), *which is assumed to exist throughout this article.* The dimension d of $\mathcal{S}_{Y|\mathbf{X}}$ is called the structural dimension of the regression, and we use a $p \times d$ matrix $\boldsymbol{\gamma}$ to denote a basis in $\mathcal{S}_{Y|\mathbf{X}}$.

The CS represents the minimal subspace that preserves the original information relative to the regression, in the sense that the conditional distribution of $Y|P_{\boldsymbol{\gamma}}\mathbf{X}$ is the same as that of $Y|\mathbf{X}$. The CS has several useful properties, among which is predictable change under full rank affine transformations of the predictor: if $\mathbf{a} \in R^p$, and $\mathbf{A} : R^p \to R^p$ is a full rank linear operator, then

$$\mathcal{S}_{Y|\mathbf{AX}+\mathbf{a}} = (\mathbf{A}^T)^{-1}\mathcal{S}_{Y|\mathbf{X}}.$$

Consequently, the dimensions of these two subspaces are the same and one can be obtained from the other. For additional background on these ideas, see Cook (1998), Cook and Weisberg (1999), and Chiaromonte and Cook (2002).

The central mean subspace (Cook and Li, 2002) is a subspace of the CS designed to characterize just the mean function $\mathrm{E}(Y|\mathbf{X})$. In this article we are concerned only with the CS and thus estimation methods like pHd (Li, 1992), IHT (Cook and Li, 2002), OPG and MAVE (Xia, et al., 2002), as well as derivative methods (Härdle and Stoker, 1989; Samarov, 1993), single-index models (Ichimura, 1993; Hristache, Juditsky and Spokoiny, 2001), and adaptive methods (Hristache, Juditsky, Polzehl and Spokoiny, 2001) that target the central mean subspace are not a focus of this article.

1.2 CS Estimation Methods and Their Conditions

CS estimation methods like IRE (Cook and Ni, 2005) and its ancestor SIR (Li, 1991) require that $\mathrm{E}(\mathbf{X}|\boldsymbol{\gamma}^T\mathbf{X} = \boldsymbol{\nu})$ be a linear function of $\boldsymbol{\nu}$. Partitioning the range of a continuous response into intervals H_s, $s = 1, \ldots, h$, this *linearity condition* implies that $\boldsymbol{\xi}_s \equiv \mathrm{Var}(\mathbf{X})^{-1}\{\mathrm{E}(\mathbf{X}|Y \in H_s) - \mathrm{E}(\mathbf{X})\} \in \mathcal{S}_{Y|\mathbf{X}}$. With $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_h)$, we then have $\mathcal{S}(\boldsymbol{\xi}) \subseteq \mathcal{S}_{Y|\mathbf{X}}$, which is the basis for the methodology. In addition to the linearity condition, the *coverage condition* – $\mathcal{S}(\boldsymbol{\xi}) = \mathcal{S}_{Y|\mathbf{X}}$ – is often imposed to insure that the CS can be estimated fully.

The linearity condition is typically easier than the coverage condition to address in practice. It holds for elliptically contoured predictors, to a good approximation when p is large relative to d (Hall and Li, 1993), and can often be induced to an adequate approximation by re-weighting (Cook and Nachtsheim, 1994) or transforming the predictors. There are no such methods or diagnostics for the coverage condition, and consequently it is often dismissed by assumption, although it is known to fail in simple symmetric models like $Y = X_1^2 + \varepsilon$, where X_1 and ε are independent standard normal random variables (Cook and Weisberg, 1991). Models in which the coverage condition fails must be carefully crafted and therefore might be judged to be rare in practice, but methods relying on the coverage condition will have low power for "nearby" models like $Y = (\mu + X_1)^2 + \varepsilon$ when $|\mu|$ is small. SAVE (Cook and Weisberg, 1991) is a second order method that mitigates the coverage condition, but it requires additional constraints on

the marginal distribution of the predictors, and a relatively large number of observations in each H_s.

Ai (1997) considered a model with less restrictive distributions on the predictor vector, which is in fact equivalent to the CS in (1). However, Ai's method involves high-dimensional density estimation. Zhu and Zeng (2006) proposed a method that targets the central mean subspace of the characteristic function of the conditional density of $Y|\mathbf{X}$. Their method recovers the CS, but they reported results only for normal predictors. A recent work of Xia (2007), which combines the kernel conditional density estimation with OPG and MAVE, also recovers the CS. Xia's method differs from ours in that it treats the predictor $\mathbf{X}$ jointly, and therefore involves a multivariate kernel estimate for the density of $\mathbf{X}$. In comparison, a key feature of our method is treating one predictor at a time, thus avoiding multivariate kernel estimation of the predictor density altogether.

1.3 A New Proposal for Estimating the CS

The new method developed in this article is based on extending to multiple-index regressions ($d > 1$) the information theory approach developed by Yin and Cook (2005) for single-index regressions ($d = 1$). A theoretical extension of this approach is conceptually straightforward, requiring neither linearity nor coverage, similarly to Ai's method (1997). However, Ai (1997) uses high-dimensional kernel density estimation. Because the convergence rate of a kernel density estimator decreases exponentially with the dimension of the kernel, whenever possible it is always preferable to use a lower dimensional kernel to a higher dimensional kernel. For this reason we break down estimation of the CS into successive single-index estimation problems. This fundamentally new aspect of our extension significantly reduces the computational complexity and, because the nonparametric procedure involves only a one-dimensional search at each stage, our method involves kernel density estimator of dimension at most 2, thus avoiding the sparsity caused by high-dimensional kernel smoothing. These gains come at a price: We still require elliptical predictors to insure the linearity condition, but the coverage condition is replaced with a relatively weak condition called *directional identifiability* that should typically hold in practice. Additionally, a continuous response does not need to be partitioned. On balance, we view the proposed method as a viable alternative to existing methods that, in effect, may require only the linearity condition in practice.

2 The Principle of Information Extraction

Unless explicitly indicated otherwise, we assume that Y has a density with respect to some (σ-finite) measure μ on $\mathbb{R}$, and $\mathbf{X}$ has a density with respect to the Lebesgue measure λ, and $(\mathbf{X}, Y)$ has a joint density with respect to $\mu \times \lambda$. For convenience, we

denote the densities of $\mathbf{X}$, Y, $Y|\mathbf{X}$ by $p(\mathbf{X})$, $p(Y)$, and $p(Y|\mathbf{X})$, and so on, keeping in mind that the symbol p in each case indicates a different function. Because densities p will always appear together with their arguments in our exposition, this abbreviation should cause no ambiguity.

Let $\mathbf{h} \in \mathbb{R}^{p\times k}$, $k \leq p$. We define the information index $\mathcal{I}(\mathbf{h})$ for dimension reduction as any of the following equivalent forms:

$$\mathcal{I}(\mathbf{h}) \equiv \mathrm{E}\left[\log \frac{p(\mathbf{h}^T\mathbf{X}|Y)}{p(\mathbf{h}^T\mathbf{X})}\right] = \mathrm{E}\left[\log \frac{p(\mathbf{h}^T\mathbf{X}, Y)}{p(Y)p(\mathbf{h}^T\mathbf{X})}\right] = \mathrm{E}\left[\log \frac{p(Y|\mathbf{h}^T\mathbf{X})}{p(Y)}\right].$$

These forms correspond to different interpretations: the first reflects inverse regression; the second is the informational correlation, and the third, ignoring a constant, is the expected log-likelihood for forward regression. The basic idea behind our proposal is to maximize sample versions of $\mathcal{I}$ in an effort to estimate a basis for the CS, using density estimates in place of the densities. In short, we seek estimates of the CS of the form $\mathcal{S}\{\arg\max \hat{\mathcal{I}}(\mathbf{h})\}$, where $\hat{\mathcal{I}}$ indicates an estimator of $\mathcal{I}$, which replaces the expectations in $\mathcal{I}$ with sample averages and replaces the densities with their nonparametric estimators.

The information index $\mathcal{I}$ is an adaptation of the Kullback-Leibler information to the dimension reduction context. As such it is connected nicely to various notions in dimension reduction, such as sufficient and maximal reduction, echoing the relation between the Kullback-Leibler information to such classical notions as sufficient and minimal sufficient statistics. The following properties are direct extensions of the single-index case ($d = 1$) by Yin and Cook (2005).

Proposition 1 *Let $\mathbf{h}_j$ be a $p \times k_j$ matrix, $j = 1, 2$, and let $\mathbf{h}$ be a $p \times k$ matrix.*

1. *If $\mathcal{S}(\mathbf{h}) \subseteq \mathcal{S}(\mathbf{h}_1)$, then $\mathcal{I}(\mathbf{h}) \leq \mathcal{I}(\mathbf{h}_1)$. If $\mathcal{S}(\mathbf{h}) = \mathcal{S}(\mathbf{h}_1)$, then $\mathcal{I}(\mathbf{h}) = \mathcal{I}(\mathbf{h}_1)$.*
2. *$\mathcal{I}(\mathbf{h}) = \mathcal{I}(I_p)$ if and only if $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{h}^T\mathbf{X}$.*
3. *$\mathcal{I}(\mathbf{h}) \geq 0$ for all $\mathbf{h}$, and $\mathcal{I}(\mathbf{h}) = 0$ if and only if $Y \perp\!\!\!\perp \mathbf{h}^T\mathbf{X}$.*
4. *$\mathcal{I}(\boldsymbol{\gamma}) \geq \mathcal{I}(\mathbf{h})$, with equality if and only if $\mathcal{S}(\boldsymbol{\gamma}) = \mathcal{S}(\mathbf{h})$.*
5. *If $k_1 < k_2 < d$, then $\mathcal{I}(I) = \mathcal{I}(\boldsymbol{\gamma}) > \max \mathcal{I}(\mathbf{h}_{k_2}) > \max \mathcal{I}(\mathbf{h}_{k_1})$.*

Part (1) confirms that $\mathcal{I}$ increases as the subspace becomes larger, and that it depends on $\mathbf{h}$ only through the subspace it spans. Part (2) means any dimension reduction subspace maximizes $\mathcal{I}$, and part (3) says that $Y \perp\!\!\!\perp \mathbf{X}$ if and only if $\max \mathcal{I}(\mathbf{h}) = 0$. Part (4) implies that searching through $p \times d$ matrices will yield a basis for the CS. Part (5) suggests that if one searches through successively larger subspaces then the first time the maximum is reached yields a basis for the CS. Thus we can always find the CS in the population by doing multi-dimensional searches. However, this is very difficult in practice since multi-dimensional density estimation can be quite problematic.

An alternative procedure that avoids high-dimensional density estimation is based on a sequence of *orthogonal 1D searches*. Letting $\mathbf{a}$ denote a length 1 vector, find a first direction $\mathbf{a}_1$ by maximizing $\mathcal{I}(\mathbf{a})$ over $\mathbb{R}^p$. The next vector, $\mathbf{a}_2$, will then be found by maximizing $\mathcal{I}(\mathbf{a})$ subject to the constraint $\mathbf{a}^T\mathbf{a}_1 = 0$. Similarly, $\mathbf{a}_k$, $k = 3, \ldots, d$ are defined as

$$\mathbf{a}_k = \text{argmax}\{\mathcal{I}(\mathbf{a}) : \mathbf{a} \perp \mathcal{S}(\mathbf{a}_1, \ldots, \mathbf{a}_{k-1})\}.$$

This procedure is like determining the eigenvectors of a square matrix $\mathbf{A}$ by doing orthogonal 1D searches to maximize the quadratic form $\mathbf{a}^T\mathbf{A}\mathbf{a}$. Thus we always search one dimension at a time with the number of estimated parameters in the kth direction being $p - k + 1$, avoiding multi-dimensional search at any time. Moreover, under conditions given in the next section, the successive maximization of $\mathcal{I}(\mathbf{a})$ is in fact *equivalent* to the joint maximization of $\mathcal{I}(\mathbf{a}_1, \ldots, \mathbf{a}_k)$, in the sense that it still recovers the CS fully, without loss of any regression information. For convenience of exposition and without loss of generality we work in the scale of the standardized predictor $\mathbf{Z} = \text{Var}(\mathbf{X})^{-1/2}(\mathbf{X} - \text{E}(\mathbf{X}))$, and let the columns of the $p \times d$ matrix $\boldsymbol{\beta}$ be a basis for $\mathcal{S}_{Y|\mathbf{Z}}$, the CS in the $\mathbf{Z}$ scale. Recall that $\boldsymbol{\gamma}$ is a basis for the CS in the $\mathbf{X}$ scale.

3 Successive Direction Extraction

Let $\mathbf{a}_1, \ldots, \mathbf{a}_d$ be the d successive maximizers of the information $\mathcal{I}$ described in the last section. We will say that the successive maximization is *Fisher consistent* (see, for example, Cox and Hinkley, 1974) if these vectors span the CS. If $d = 0$ no maximization is required and we say that the procedure is Fisher consistent by definition. As previously demonstrated by Yin and Cook (2005), the procedure is Fisher consistent when $d = 1$. If $d = p$ the procedure is also Fisher consistent because $(\mathbf{a}_1, \ldots, \mathbf{a}_p)$ must span $\mathcal{S}_{Y|\mathbf{X}} = \mathbb{R}^p$. Thus an issue arises only when $2 \leq d < p$. Following the foundations set in Sections 3.1 and 3.2, we establish Fisher consistency in Section 3.3. The significance of this is two fold: it substantially reduces the computation cost and it avoids the use of high dimensional kernels. At the sample level, we maximize the estimated version of $\mathcal{I}$ to estimate the CS. This is analogous to maximizing a likelihood to estimate a parameter in a parametric setting (maximum likelihood estimation). But there are important differences — the "parameter" in our context is a direction, or more generally a subspace, rather than a point, and the "likelihood" itself is non-parametrically estimated. Maximum likelihood estimation is based on a fundamental assumption, the identifiability of the parameter, which guarantees the uniqueness of the maximizer of the expected log likelihood. Here, we need a similar assumption. However, because the objective of estimation here is a subspace, the classical definition of identifiability no longer suits our purpose. So we first adapt it to the new estimation problem of dimension reduction.

3.1 Directional Identifiability

The classical concept of identifiability can be stated as follows. Let $\mathcal{N} = \{\nu_\theta : \theta \in \Theta\}$ be a parametric family of probability measures. We say that $\mathcal{N}$ is identifiable if, for any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$, we have $\nu_{\theta_1} \neq \nu_{\theta_2}$. Equivalently, we can define identifiability via the following scheme. We say that a point θ in the parameter space Θ is identifiable if, for any $\theta' \neq \theta$, $\theta' \in \Theta$, $\nu_\theta \neq \nu_{\theta'}$. We say that the parametric family $\mathcal{N}$ is identifiable if every point in Θ is identifiable. Our definition of directional identifiability will resemble the second statement.

To understand the need for, and the meaning of, this new concept we start with a special case. Note that the very possibility of dimension reduction implies some directions are irrelevant. Naturally, we do not expect these directions to be identifiable. In the next definition $\mathcal{S}$ is a one-dimensional subspace, which we call a direction in $\mathbb{R}^p$.

Definition 1 *When $0 < d < p$, a direction $\mathcal{S}$ in $\mathbb{R}^p$ is an identifiable direction if, for any other direction $\mathcal{S}'$, with $\mathcal{S}' \neq \mathcal{S}$, we have*

$$\Pr\{p(Y|P_{\mathcal{S}}\mathbf{Z}) \neq p(Y|P_{\mathcal{S}'}\mathbf{Z})\} > 0.$$

The conditional density $p(Y|\mathbf{Z})$ is directionally identifiable to the first order if every direction in the CS is an identifiable direction.

This definition differs from the classical identifiability in two aspects: only subspaces, or directions, are relevant in the definition, and, for the conditional density $p(Y|\mathbf{Z})$ to be directionally identifiable, we require only those directions in the CS to be identifiable. In fact we will see later that directions outside the CS may well be unidentifiable. Roughly, this definition states that each direction in $\mathcal{S}_{Y|\mathbf{Z}}$ determines a unique conditional density. Associating a conditional density $p(Y|P_{\mathcal{S}}\mathbf{Z})$ with a 2D plot of Y versus $\mathbf{a}^T\mathbf{Z}$, where $\mathbf{a}$ is a vector that spans $\mathcal{S}$, the directional identifiability of $p(Y|\mathbf{Z})$ in Definition 1 can also be interpreted informally as requiring that all 2D plots for directions $\mathcal{S} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ must be stochastically distinct. We now generalize this definition to kth order directional identifiability. Although we will be searching for one direction at a time, we may still "pass through" subspaces of various dimensions while building up to the CS. For this reason the notion of kth order identifiability is relevant.

Definition 2 *Assume $0 < k \leq d < p$. A k-dimensional subspace $\mathcal{S}$ in $\mathbb{R}^p$ is identifiable if, for any other k-dimensional subspace $\mathcal{S}'$ with $\mathcal{S} \neq \mathcal{S}'$, we have*

$$\Pr\{p(Y|P_{\mathcal{S}}\mathbf{Z}) \neq p(Y|P_{\mathcal{S}'}\mathbf{Z})\} > 0.$$

If every k-dimensional subspace of $\mathcal{S}_{Y|\mathbf{Z}}$ is identifiable then we say that $p(Y|\mathbf{Z})$ is directionally identifiable to the kth order.

To further understand directional identifiability let us examine some special subspaces that are *not* identifiable. Suppose $0 < d < k \leq p$. Then every two subspaces in $\mathbb{R}^p$ containing the CS give identical conditional distribution. So every subspace of dimension larger than d that contains $\mathcal{S}_{Y|\mathbf{Z}}$ is un-identifiable. The CS is defined as the intersection of all dimension reduction subspaces and consequently it must be an identifiable subspace. This is because, otherwise, there is a different dimension reduction subspace of the same dimension, and the CS cannot be the intersection of all dimension reduction subspaces. Finally, sometimes every subspace of the orthogonal complement of the CS is un-identifiable. By Proposition 8 (stated later), if Q is the orthogonal projection onto $\mathcal{S}^{\perp}_{Y|\mathbf{Z}}$, and if $\mathbf{Z}$ or $\mathbf{Z}|Y$ is normal, then $Y \perp\!\!\!\perp Q\mathbf{Z}$. In this case every subspace in $\mathcal{S}(Q)$ gives rise to the same conditional distribution, which is the unconditional density of Y. Hence, in this case, no subspace in $\mathcal{S}(Q)$ is identifiable. We summarize these properties in the next proposition.

Proposition 2 *Assume $0 < d < p$. The following hold regarding directional identifiability:*

1. *No subspace of $\mathbb{R}^p$ with dimension larger than d that contains the CS is identifiable.*
2. *The CS is an identifiable subspace of $\mathbb{R}^p$.*
3. *If $Y \perp\!\!\!\perp Q\mathbf{Z}$, then no proper subspace of $\mathcal{S}^{\perp}_{Y|\mathbf{Z}}$ is identifiable.*

These properties further indicate that, to define the directional identifiability of $p(Y|\mathbf{Z})$, we should require only the subspaces in the CS to be identifiable, and should define it only up to the order d:

Definition 3 *We say that $p(Y|\mathbf{Z})$ is directionally identifiable if any of the following three conditions holds: (a) $d = 0$, (b) $d = 1$ or (c) $2 \leq d \leq p$ and $p(Y|\mathbf{Z})$ is directionally identifiable to orders $1, 2, \ldots, d-1$.*

Regressions with $d = 0$ are defined to be directionally identifiable just to include the trivial case. Regressions with $d = 1$ are defined to be so based on point 2 of Proposition 2. For the same reason we require directionally identifiability only up to order $d-1$ when $d \geq 2$.

3.2 Sufficient Conditions for Directional Identifiability

Definition 3 indicates that to establish directional identifiability of $p(Y|\mathbf{Z})$ we may have to establish k-th order identifiability for several values of k. In this subsection we will demonstrate that directional identifiability of $p(Y|\mathbf{Z})$ is in fact equivalent to the simpler condition of the first-order directional identifiability of $p(Y|\mathbf{Z})$ if a mild condition is satisfied. This mild condition has to do with what we call "M-sets" (M for "matching"),

which, together with its implications Proposition 3 and Corollary 1, will also play an important role in establishing Fisher consistency in the next few subsections.

Definition 4 *A set A in $\mathbb{R}^s \times \mathbb{R}^t$ is called an M-set if, for every two pairs $(\mathbf{x}_1, \mathbf{x}_2)$ and $(\mathbf{x}_1', \mathbf{x}_2')$ in A, there is a set of pairs $(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}), \ldots, (\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)})$ with $(\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)})$ denoting $(\mathbf{x}_1, \mathbf{x}_2)$ and $(\mathbf{x}_1^{(m)}, \mathbf{x}_2^{(m)})$ denoting $(\mathbf{x}_1', \mathbf{x}_2')$ such that (1) $(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}) \in A$, $i = 1, \ldots, m$; (2) for each $i = 1, \ldots, m-1$, at least one of the following is true: $\mathbf{x}_1^{(i)} = \mathbf{x}_1^{(i+1)}$ or $\mathbf{x}_2^{(i)} = \mathbf{x}_2^{(i+1)}$.*

Intuitively, a set in $\mathbb{R} \times \mathbb{R}$ is an M-set if every two points in it can be connected by a "stairway", each of whose corner points belong to the set. The stairs are allowed to have different sizes and to head in different directions. Moreover, only the corners — not the whole stairway — are required to be in the set. This is a very general condition. For example, it is easy to see that any set in $\mathbb{R}^s \times \mathbb{R}^t$ that is connected and open is an M-set, so any convex and open set in $\mathbb{R}^s \times \mathbb{R}^t$ is an M-set. In fact, it is easy to construct an M-set that is not even connected. The next proposition and its corollary do not require densities and so are stated in terms of distribution functions $F_{(\cdot)}$.

Proposition 3 *Let Y, $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{U}_3$ be random elements. Let Ω be the sample space of $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ and, for each fixed $\mathbf{u}_3$, let $\Omega_{12}(\mathbf{u}_3) = \{(\mathbf{u}_1, \mathbf{u}_2) : (\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3) \in \Omega\}$. Suppose*

1. *$F_{Y|\mathbf{U}_1\mathbf{U}_3} = F_{Y|\mathbf{U}_2\mathbf{U}_3}$.*
2. *For each $\mathbf{u}_3$, $\Omega_{12}(\mathbf{u}_3)$ is an M-set.*

Then $Y \perp\!\!\!\perp \mathbf{U}_1 | \mathbf{U}_3$ and $Y \perp\!\!\!\perp \mathbf{U}_2 | \mathbf{U}_3$.

An important special case of this proposition is when $\mathbf{U}_3$ is a constant, which will also be useful for our exposition. We record it below as a corollary.

Corollary 1 *Let Y, $\mathbf{U}_1$, $\mathbf{U}_2$ be random elements. Let Ω be the sample space of $(\mathbf{U}_1, \mathbf{U}_2)$. Suppose that $F_{Y|\mathbf{U}_1} = F_{Y|\mathbf{U}_2}$ and that Ω is a M-set. Then $Y \perp\!\!\!\perp \mathbf{U}_1$ and $Y \perp\!\!\!\perp \mathbf{U}_2$.*

Returning to the main theme and again assuming densities, consider characteristics of regressions in which $p(Y|\mathbf{Z})$ is not directionally identifiable so that we must have $d \geq 2$. Assume that $p(Y|\mathbf{Z})$ is not identifiable to first order and thus it is not directionally identifiable. Then there are two distinct directions $\mathcal{S} \subset \mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}' \subset \mathbb{R}^p$ such that $p(Y|P_{\mathcal{S}}\mathbf{Z}) = p(Y|P_{\mathcal{S}'}\mathbf{Z})$. Assuming that the sample space of $(P_{\mathcal{S}}\mathbf{Z}, P_{\mathcal{S}'}\mathbf{Z})$ is an M-set, it follows from Corollary 1 that $Y \perp\!\!\!\perp P_{\mathcal{S}}\mathbf{Z}$ and that $Y \perp\!\!\!\perp P_{\mathcal{S}'}\mathbf{Z}$. Thus, while $p(Y|\mathbf{Z}) = p(Y|P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{Z})$ represents the unique minimal reduction, the response must

be independent of at least one projected direction within the CS for first order identifiability to fail, $p(Y|P_{\mathcal{S}}P_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{Z}) = p(Y|P_{\mathcal{S}}\mathbf{Z}) = p(Y)$. As a specific instance of this failure, let (X_1, X_2) be uniformly distributed on the unit disk, $X_1^2 + X_2^2 < 1$, and let $(Y, X_3)|(X_1, X_2)$ follow a bivariate normal distribution with means 0, variances 1 and covariance X_1^2. Then $\mathrm{E}(Y|\mathbf{X}) = X_1^2 X_3$ and so $d = 2$. However, $Y \perp\!\!\!\perp (X_1, X_2)$ and thus $p(Y|\mathbf{X})$ is not directionally identifiable.

The next proposition, which gives sufficient conditions for first order directional identifiability to imply directional identifiability, is a logical consequence of our discussion:

Proposition 4 *If $d \geq 3$, assume that the sample space of $(\mathbf{A}_1^T\mathbf{Z}, \mathbf{A}_2^T\mathbf{Z})$ is an M-set for any $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{p\times k}$, $k = 1, \ldots, d-1$. Then $p(Y|\mathbf{Z})$ is directionally identifiable if (a) $d = 0$, if (b) $d = 1$ or if (c) $d \geq 2$ and $p(Y|\mathbf{Z})$ is first-order directionally identifiable.*

Remark To conclude this section, it is worth mentioning that the concept of an M-set crystallizes the sufficient conditions for the existence of the CS, though this has no direct bearing on the main result of this article. Indeed, as can be seen from Proposition 6.4 of Cook (1998), the M-set assumption is sufficient for the proof of this existence, though much stronger conditions were assumed there. The proof of the following proposition is similar to that of Proposition 6.4 of Cook (1998), and is omitted.

Proposition 5 *Suppose that $\boldsymbol{\beta}_i$, $i = 1, 2$, are $p\times q_i$ matrices with $q_i \leq p$ such that $\mathcal{S}(\boldsymbol{\beta}_1)$ and $\mathcal{S}(\boldsymbol{\beta}_2)$ are both dimension reduction subspaces. Let $\boldsymbol{\beta}_3$ be a $p\times q_3$ dimensional matrix whose columns span the subspace $\mathcal{S}(\boldsymbol{\beta}_1) \cap \mathcal{S}(\boldsymbol{\beta}_2)$. Suppose, in addition, that for each $\mathbf{u} \in \mathbb{R}^{q_3}$ such that $\boldsymbol{\beta}_3^T\mathbf{z} = u$ for some $\mathbf{z} \in \Omega_{\mathbf{Z}}$, the set*

$$\left\{ \left(\boldsymbol{\beta}_1^T\mathbf{z}, \boldsymbol{\beta}_2^T\mathbf{z}\right) : \boldsymbol{\beta}_3^T\mathbf{z} = \mathbf{u}, \mathbf{z} \in \Omega_{\mathbf{Z}} \right\} \tag{2}$$

is an M-set in $\mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$. Then $\mathcal{S}(\boldsymbol{\beta}_1) \cap \mathcal{S}(\boldsymbol{\beta}_2)$ is a also dimension reduction subspace.

Consequently, under the assumptions of this proposition, the CS exists.

3.3 Sufficient Conditions for Fisher Consistency

In this subsection we establish Fisher consistency of successive maximization of $\mathcal{I}$ under a set of sufficient conditions, which are to be verified under three different circumstances in the next few subsections.

Let $\mathcal{S}$ be a proper subspace of $\mathcal{S}_{Y|\mathbf{Z}}$. Then $\mathcal{S}_{Y|Z}^{\perp}$ is a proper subset of $\mathcal{S}^{\perp}$, and the latter has the decomposition

$$\mathcal{S}^{\perp} = \left(\mathcal{S}^{\perp} \cap \mathcal{S}_{Y|Z}\right) \oplus \left(\mathcal{S}^{\perp} \cap \mathcal{S}_{Y|Z}^{\perp}\right).$$

It is evident that the successive maximization of $\mathcal{I}$ will be Fisher consistent if, for any proper subspace $\mathcal{S}$ of $\mathcal{S}_{Y|\mathbf{Z}}$, the maximizer of $\mathcal{I}$ in $\mathcal{S}^\perp$ belongs to $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$. If this is true, then, in the first step, we can take $\mathcal{S}$ to be $\varnothing$, and $\mathbf{a}_1$ must be in $\mathcal{S}_{Y|\mathbf{Z}}$; and if $\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}$ are in $\mathcal{S}_{Y|\mathbf{Z}}$, then, in the the kth step, we can take $\mathcal{S}$ to be the subspace spanned by $\mathbf{a}_1, \ldots, \mathbf{a}_{k-1}$, and $\mathbf{a}_k$ must be in $\mathcal{S}_{Y|\mathbf{Z}}$. Thus $\mathbf{a}_1, \ldots, \mathbf{a}_d$ must all belong to $\mathcal{S}_{Y|\mathbf{Z}}$.

Let P be the orthogonal projection onto $\mathcal{S}_{Y|\mathbf{Z}}$ and $Q = I_p - P$ be its orthogonal complement in $\mathbb{R}^p$. Similarly, let P_1 be the orthogonal projection onto $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ and Q_1 be the projection onto $\mathcal{S}^\perp \cap \mathcal{S}^\perp_{Y|\mathbf{Z}}$. Note that, because $\mathcal{S}$ is a proper subset of $\mathcal{S}_{Y|\mathbf{Z}}$ we have $\mathcal{S}^\perp \cap \mathcal{S}^\perp_{Y|\mathbf{Z}} = \mathcal{S}^\perp_{Y|\mathbf{Z}}$ and hence $Q_1 = Q$.

Proposition 6 *Suppose $\mathcal{S}$ and P_1 are as defined in the last two paragraphs. Suppose, furthermore, that*

1. *the density $p(Y|\mathbf{Z})$ is directionally identifiable to the first order,*
2. *for any $\mathbf{a} \in \mathcal{S}^\perp$, we have*

$$Y \perp\!\!\!\perp \mathbf{a}^T\mathbf{Z} | (P_1\mathbf{a})^T\mathbf{Z}. \tag{3}$$

Then, for any vector $\mathbf{a}$ that is in $\mathcal{S}^\perp$ but not in $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$, there is a non-zero vector $\mathbf{b}$ in $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ such that $\mathcal{I}(\mathbf{b}) > \mathcal{I}(\mathbf{a})$.

Proposition 6 has no content when $d = 0$ or $d = p$ because in those cases it is not possible to satisfy its premise. When $d = 1$ the only case that satisfies the Proposition's premise is $\mathcal{S}^\perp = \mathbb{R}^p$, and then it reduces to essentially a restatement of point 4 of Proposition 1. This proposition then becomes important when $2 \le d < p$, which covers exactly the cases needed from the discussion at the opening of Section 3.

Proposition 6 implies that the maximizer of $\mathcal{I}$ in $\mathcal{S}^\perp$ must be in $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ which, combined with the argument preceding Proposition 6 implies that the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_d$ must all belong to $\mathcal{S}_{Y|\mathbf{Z}}$. We record this conclusion along with the discussion at the opening of Section 3 in the next corollary.

Corollary 2 *The successive maxima $\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ of $\mathcal{I}$ are Fisher consistent if $d = 0, 1$ or p, and are otherwise so under the assumptions of Proposition 6.*

3.4 Predictors with Elliptical Distributions

The key assumption for Proposition 6, and hence the Fisher consistency of successive maximization of $\mathcal{I}$, is the conditional independence condition (3). This condition holds in any one of three situations: (a) $\mathbf{Z}$ has an elliptical density and $p - d \ge 2$, plus a very mild additional assumption, (b) $\mathbf{Z}$ is normal, and (c) $\mathbf{Z}|Y$ is normal. Among these, (a)

is the most generally applicable, because essentially it requires only ellipticity of the density of $\mathbf{Z}$ ($p-d \geq 2$ is met in most applications); (b) and (c) do not require $p-d \geq 2$ but impose more restrictive structure on $\mathbf{Z}$.

Using M-sets, we now establish the conditional independence (3) under the first set of assumptions (a). In establishing (3), the random variables with which we will be concerned are of the form $\mathbf{a}^T\mathbf{Z}$. The next lemma specializes M-sets to this context. For simplicity we have made the assumptions stronger than needed, but even so the result is sufficiently general for our purpose.

Lemma 1 *Suppose that C is an open and convex set in $\mathbb{R}^p$, and that $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are linearly independent vectors in $\mathbb{R}^p$. Then the following assertions hold:*

1. *The set $\{(\mathbf{a}_1^T\mathbf{z}, \mathbf{a}_2^T\mathbf{z}) : \mathbf{z} \in C\}$ is an M-set in $\mathbb{R} \times \mathbb{R}$.*
2. *For each $u \in \mathbb{R}$ such that $\mathbf{a}_3^T\mathbf{z} = u$ for some $\mathbf{z} \in C$, the set $\{(\mathbf{a}_1^T\mathbf{z}, \mathbf{a}_2^T\mathbf{z}) : \mathbf{a}_3^T\mathbf{z} = u, \mathbf{z} \in C\}$ is an M-set in $\mathbb{R} \times \mathbb{R}$.*

The proof of this lemma is straightforward and omitted. But it may be intuitively clear that the sets in both parts of the lemma are open and connected, and hence are M-sets.

We now establish the conditional independence (3) using Proposition 3 and its corollary. Let $(\mathbf{u}_1, \ldots, \mathbf{u}_d)$ be an orthonormal basis of $\mathcal{S}_{Y|\mathbf{Z}}$ and $(\mathbf{w}_1, \ldots, \mathbf{w}_{p-d})$ be an orthonormal basis in $\mathcal{S}_{Y|\mathbf{Z}}^{\perp}$. Let $\mathbf{U}$ denote the matrix $(\mathbf{u}_1, \ldots, \mathbf{u}_d)$ and let $\mathbf{W}$ denote the matrix $(\mathbf{w}_1, \ldots, \mathbf{w}_{p-d})$. Let $\mathbf{A}$ be a $(p-d) \times (p-d)$ orthogonal matrix, and let

$$\mathbf{B} = \mathbf{U}\mathbf{U}^T + \mathbf{W}\mathbf{A}\mathbf{W}^T. \tag{4}$$

This matrix rotates the Q-component of a vector while leaving the P-component of the vector intact. The next two lemmas concern the properties of $\mathbf{A}$ and $\mathbf{B}$. The proof of Lemma 2 is straightforward and omitted.

Lemma 2 *The matrix $\mathbf{B}$ is a $p \times p$ orthogonal matrix. Moreover, if $\mathcal{S}$, P_1 and Q_1 are as defined in Section 3.3, then*

$$P\mathbf{B} = \mathbf{B}P = P, \quad P_1\mathbf{B} = \mathbf{B}P_1 = P_1, \quad Q_1\mathbf{B} = \mathbf{B}Q_1 = \mathbf{W}\mathbf{A}\mathbf{W}^T.$$

In the following $\mathcal{S}$, P_1 and Q_1 are as defined in Section 3.3.

Lemma 3 *Suppose that $p - d \geq 2$, and that $\mathbf{a}$ is a vector in $\mathbb{R}^p$ but not in $\mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$.*

1. *If $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}^{\perp}$ then there is an $(p-d) \times (p-d)$ orthogonal matrix $\mathbf{A}$ such that $\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a}$, $P_1\mathbf{a}$, and $Q_1\mathbf{a}$ are linearly independent.*

2. *If* $\mathbf{a} \in \mathcal{S}^{\perp} \cap \mathcal{S}^{\perp}_{Y|\mathbf{Z}}$ *then there is a* $(p-d) \times (p-d)$ *orthogonal matrix* $\mathbf{A}$ *such that* $\mathbf{WAW}^T\mathbf{a}$ *and* $\mathbf{a}$ *are linearly independent.*

We are now ready to present our result.

Proposition 7 *Suppose that* $p - d \geq 2$ *and that* $\mathbf{Z}$ *has an elliptical density. Suppose that the support* $\Omega_{\mathbf{Z}}$ *of* $\mathbf{Z}$ *is convex and satisfies* $\Pr(\mathbf{Z} \in \Omega^0_{\mathbf{Z}}) = 1$*, where* $\Omega^0_{\mathbf{Z}}$ *is the interior of* $\Omega_{\mathbf{Z}}$*. Then (3) holds for any* $\mathbf{a} \in \mathcal{S}^{\perp}$*.*

3.5 Predictors with normal or conditional normal distributions

In this section we establish the conditional independence (3) under the assumption that either $\mathbf{Z}$ or $\mathbf{Z}|Y$ is normally distributed. Proposition 7 established (3) for normal $\mathbf{Z}$ when $p - d \geq 2$, but here we do not impose this condition. The development builds upon some population properties of the SIR (Li, 1991) and SAVE (Cook and Weisberg, 1991) estimators. Still working with the standardized predictor $\mathbf{Z}$, let

$$\mathbf{M}_{\mathrm{SIR}} = \mathrm{Var}\left[\mathrm{E}(\mathbf{Z}|Y)\right], \quad \text{and} \quad \mathbf{M}_{\mathrm{SAVE}} = \mathrm{E}\left[I_p - \mathrm{Var}(\mathbf{Z}|Y)\right]^2.$$

Recall that Ω_Y is the sample space of Y.

Lemma 4 *The following properties hold for* $\mathbf{M}_{\mathrm{SAVE}}$ *and* $\mathbf{M}_{\mathrm{SIR}}$*:*

1. *There is a set* $A \subseteq \Omega_Y$ *with* $\Pr(Y \in A) = 1$*, such that*

$$\mathcal{S}(\mathbf{M}_{\mathrm{SAVE}}) = \mathrm{span}\{I_p - \mathrm{Var}(\mathbf{Z}|Y = y) : y \in A\}.$$

   And (consequently), $\mathbf{M}_{\mathrm{SIR}} \subseteq \mathbf{M}_{\mathrm{SAVE}}$*.*

2. *If* $\mathbf{Z}|Y$ *follows a multivariate normal distribution, then* $\mathrm{E}(\mathbf{Z}|P\mathbf{Z}) = P\mathbf{Z}$ *and* $\mathrm{Var}(\mathbf{Z}|P\mathbf{Z}) = Q$*.*

3. *If* $\mathbf{Z}|Y$ *follows a multivariate normal distribution, then* $\mathcal{S}(\mathbf{M}_{\mathrm{SAVE}}) = \mathcal{S}_{Y|\mathbf{Z}}$*. If, moreover,* $\mathrm{Var}(\mathbf{Z}|Y)$ *is nonrandom matrix, then* $\mathbf{M}_{\mathrm{SIR}} = \mathbf{M}_{\mathrm{SAVE}} = \mathcal{S}_{Y|\mathbf{Z}}$*.*

Part 2 of the lemma extends a similar result of Cook and Lee (1999), and part 3 is an extension of some results in Cook and Lee (1999) and Cook and Yin (2001). The proof of this lemma will be omitted. We are now ready to state our result.

Proposition 8 *Suppose that either* $\mathbf{Z}$ *or* $\mathbf{Z}|Y$ *follows a* p*-dimensional multivariate normal distribution. Then*

$$Q\mathbf{Z} \perp\!\!\!\perp (Y, P\mathbf{Z}). \tag{5}$$

And consequently, (3) holds.

Thus, we have established the Fisher consistency of successive maximization of information under any one of the three sets of conditions. For clarity, we summarize the results in Sections 3.3, 3.4, and 3.5 into the next corollary.

Corollary 3 *Suppose that the conditional density $p(Y|\mathbf{Z})$ is directionally identifiable to the first order. Suppose any one of the following conditions holds:*

1. *the standardized predictor $\mathbf{Z}$ has an elliptical density, and has a convex support $\Omega_{\mathbf{Z}}$ with $\Pr(\mathbf{Z} \in \Omega_{\mathbf{Z}}^0) = 1$; moreover $p - d \geq 2$;*
2. *$\mathbf{Z}$ follows a p-dimensional multivariate normal distribution;*
3. *$\mathbf{Z}|Y$ follows a p-dimensional multivariate normal distribution.*

Then the successive 1-D maximizers of $\mathcal{I}$, $\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$ (as defined in Section 3.3), is Fisher consistent.

3.6 Algorithm

The Fisher consistency of successive maximization means that we can always search one direction at a time for a multiple-index regression, essentially reducing it to a single-index regression. Thus we can use the algorithm suggested by Yin and Cook (2005) for single-index regressions. To summarize, we describe the algorithm at the population level. The sample version is obtained straightforwardly by substituting estimates.

1. Let $\mathbf{a}_1 = \arg\max_{\mathbf{a} \in \mathbb{R}^p} \mathrm{E}[\log p(Y|\mathbf{a}^T\mathbf{Z})]$ be the first direction obtained by our search using Yin and Cook's (2005) algorithm; that is, $\mathbf{a}_1^T\mathbf{Z}$ is our first variable. Letting $(\mathbf{a}_1, \boldsymbol{\Gamma}_1)$ be an orthogonal matrix, our second search is on $\boldsymbol{\Gamma}_1^T\mathbf{Z} \in \mathbb{R}^{p-1}$.

2. Let $\mathbf{a}_1^* \in \mathbb{R}^{p-1}$ be the direction for our second single-index search with $\boldsymbol{\Gamma}_1^T\mathbf{Z} \in \mathbb{R}^{p-1}$ as the current predictor vector; that is

$$\mathbf{a}_1^* = \arg\max \left\{ \mathrm{E}[\log p(Y|\mathbf{a}^{*T}\boldsymbol{\Gamma}_1^T\mathbf{Z})] : \mathbf{a}^* \in \mathbb{R}^{p-1} \right\}.$$

   Then our second direction in the original $\mathbf{Z}$ scale is $\mathbf{a}_2 = \boldsymbol{\Gamma}_1\mathbf{a}_1^*$.

3. Let the $p \times 1$ vectors $\mathbf{a}_1, \cdots, \mathbf{a}_{k-1}$ be the first $k-1$ directions of our search, and let $(\mathbf{a}_1, \cdots, \mathbf{a}_{k-1}, \boldsymbol{\Gamma}_{k-1})$ form an orthogonal matrix. Our next search is over the $(p-k+1) \times 1$ vector $\boldsymbol{\Gamma}_{k-1}^T\mathbf{Z}$. If the solution for our single-index search is $\mathbf{a}_{k-1}^*$, then our kth direction in the original $\mathbf{Z}$ scale is $\mathbf{a}_k = \boldsymbol{\Gamma}_{k-1}\mathbf{a}_{k-1}^*$.

4. Finally, $(\mathbf{a}_1, \cdots, \mathbf{a}_d)$ forms a basis for $\mathcal{S}_{Y|\mathbf{Z}}$. This can be linearly transformed to a basis for $\mathcal{S}_{Y|\mathbf{X}}$ as indicated previously.

4 Estimating d

In this section, we suggest a method to estimate $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$. While one can in principle determine the dimension d by carrying out a sequential asymptotic test based on the asymptotic distribution of the estimator of $\mathcal{S}_{Y|\mathbf{Z}}$, along the lines of Li (1991, 1992), Schott (1994), and Cook and Li (2004), the full asymptotic development is beyond the scope of this paper. We will leave the development of asymptotic distribution to future research.

Here we suggest an alternative permutation test. A permutation test can be used to test the independence between two random vectors. However, we are concerned with testing conditional independence, to which a permutation test cannot be directly applied. Thus we first need to convert the problem of testing conditional independence to that of testing unconditional independence. Let $\mathbf{A}_k$ be the p by k matrix $(\mathbf{a}_1, \cdots, \mathbf{a}_k)$ and let $\mathbf{A}_{p-k}$ be a p by $p-k$ matrix such that the columns of $(\mathbf{A}_k, \mathbf{A}_{p-k})$ form an orthonormal basis for $\mathbb{R}^p$. At each step of the successive maximization of the information index $\mathcal{I}$, we would like to decide if further maximization is needed. In other words, we would like to test the hypothesis that $Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}|\mathbf{A}_k^T\mathbf{Z}$. The following proposition gives the sufficient conditions under which this conditional independence is equivalent to the unconditional independence $Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}$.

Proposition 9 *Suppose that either* $\mathbf{Z}$ *or* $\mathbf{Z}|Y$ *is normally distributed. Then,*

$$Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}|\mathbf{A}_k^T\mathbf{Z} \tag{6}$$

is equivalent to $\mathbf{A}_{p-k}^T\mathbf{Z} \perp\!\!\!\perp (Y, \mathbf{A}_k^T\mathbf{Z})$. *Suppose, moreover, that* $p - d \geq 1$ *and that* $p(Y|\mathbf{Z})$ *is first-order directionally identifiable, then (6) is equivalent to* $Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}$.

Under normality or conditional normality for $\mathbf{Z}$ or $\mathbf{Z}|Y$, respectively, the ideal hypothesis is equivalent to $\mathbf{A}_{p-k}^T\mathbf{Z} \perp\!\!\!\perp (Y, \mathbf{A}_k^T\mathbf{Z})$. Neglecting bivariate deviations from this hypothesis, we propose to test the marginal implication that $Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}$ by performing a permutation test for a single direction in the regression of Y on $\mathbf{A}_{p-k}^T\mathbf{Z}$. Furthermore if $p - d \geq 1$, and directional identifiability assumption holds, then the test is exact test. Otherwise, this gives us only a lower bound on d in general, but it does avoid high-dimensional density estimation which is our goal. Testing $Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T\mathbf{Z}$ can then proceed based on part (3) of Proposition 1. The effect of any bias in estimating a density should be negligible since we use the same density estimation procedure with the same kernel and bandwidth for each permutation. Our simulations in Section 6 show that this procedure is quite effective, even for non-normal data.

The algorithm for the proposed permutation test is as follows:

1. Determine the first estimated direction $\mathbf{a}_1$, and calculate $n\hat{\mathcal{I}}(\mathbf{a}_1)$. Permute the observed predictor vectors $\{\mathbf{X}_i,\ 1 = 1, \ldots, n\}$, B times, each time searching for the best direction, $\mathbf{a}_1^b$, for $b = 1, \cdots, B$ and calculating $n\hat{\mathcal{I}}(\mathbf{a}_1^b)$. If $n\hat{\mathcal{I}}(\mathbf{a}_1)$ is less than the 5% of the cut-off point of $\{n\hat{\mathcal{I}}(\mathbf{a}_1^b), b = 1, \cdots, B\}$ then we infer $d = 0$; otherwise, the first estimated direction is $\mathbf{a}_1$.

2. Assume that the first estimated k directions are $\mathbf{A}_k = (\mathbf{a}_1, \cdots, \mathbf{a}_k)$, and again let $(\mathbf{A}_k, \mathbf{A}_{p-k})$ form an orthonormal basis for $\mathbb{R}^p$. Search for the $(k+1)$th direction, $\mathbf{a}_{k+1}$, by using data $\{(Y_i, \mathbf{A}_{p-k}^T\mathbf{Z}_i) : i = 1, \ldots, n\}$. Permute $\{\mathbf{A}_{p-k}^T\mathbf{Z}_i, 1 = 1, \ldots, n\}$ B times, each time determining the best direction and calculating $n\hat{\mathcal{I}}(\mathbf{a}_{k+1}^b)$. If $n\hat{\mathcal{I}}(\mathbf{a}_{k+1})$ is less than the 5% of the cut-off point of $n\hat{\mathcal{I}}(\mathbf{a}_{k+1}^b)$ for $b = 1, \cdots, B$, then we infer $d = k$; otherwise, the $(k+1)$th estimated direction is $\mathbf{a}_{k+1}$.

3. Update step 2, until it stops.

In practice, we take $B = 1000$.

Bootstrap test (Ye and Weiss, 2003) should also work well in this context. The main difference — apart from the difference in the re-sampling mechanisms — is that in the bootstrap test the difference between $\mathbf{A}_k$ and $\mathbf{A}_k^b$ (the bootstrap solution) is measured by their angle, whereas in permutation test this difference is measured by the difference of their information (similar to the Kullback Leibler difference), a natural outcome of our algorithm.

5 Consistency of Estimators

In this section we demonstrate the consistency of the successive estimators. Note that $\mathcal{I}(\mathbf{a})$ and $\mathrm{E}[\log p(Y|\mathbf{a}^T\mathbf{X})]$ differ only by $\mathrm{E}\log p(Y)$, which does not depend on $\mathbf{a}$. So for maximization over $\mathbf{a}$ these two quantities are equivalent. In this section we will use $\mathcal{I}(\mathbf{a})$ to denote $\mathrm{E}[\log p(Y|\mathbf{a}^T\mathbf{X})]$.

Once the first vector $\hat{\mathbf{a}}_1 = \mathrm{argmax}\{\hat{\mathcal{I}}(\mathbf{a}) : \mathbf{a} \in \mathbb{R}^p\}$ is obtained, the next vector, $\hat{\mathbf{a}}_2$, is obtained by maximizing $\hat{\mathcal{I}}(\mathbf{a})$ over the set $\{\mathbf{a} \in \mathbb{R}^p : \mathbf{a} \perp \hat{\mathbf{a}}_1\}$, which is a random set depending on $\hat{\mathbf{a}}_1$. Hence, unlike in a single index model, here we are no longer maximizing an objective function over a fixed space. The same can be said of the subsequent vectors $\hat{\mathbf{a}}_3, \ldots, \hat{\mathbf{a}}_d$. It is this aspect that we must tackle when drawing asymptotic conclusions for the multiple index model from the known asymptotic facts for the single index models.

Let $\mathbf{a} \in \mathbb{R}^p$. Because $\mathcal{I}(\mathbf{a}) = \mathcal{I}(c\mathbf{a})$ for any $c \neq 0$, whatever constraints on $\mathbf{a}$ do not change the essence of the estimation, but constraint such as $||\mathbf{a}|| = 1$ in Section 3.1 does complicate the procedure in developing consistent result due to the related derivatives. To simplify the procedure we use a different constraint. Since $\mathbf{a}_1, \ldots, \mathbf{a}_d$ are nonzero,

each of them has at least one nonzero component, and we can assume that component to be 1. For convenience, we consider only the special cases where the pth components of $\mathbf{a}_1, \ldots, \mathbf{a}_d$ are nonzero, and taken them to be 1. The general case can be treated using essentially the same argument.

To focus on the sequential nature of the result, we will derive the consistency by assuming at the outset that there is a uniformly convergent estimator $\hat{\mathcal{I}}(\mathbf{a})$ of $\mathcal{I}(\mathbf{a})$, that is,

$$\sup_{\mathbf{a}} \|\hat{\mathcal{I}}(\mathbf{a}) - \mathcal{I}(\mathbf{a})\| \xrightarrow{p} 0. \tag{7}$$

Here, the supremum is taken over a subset of $\mathbb{R}^p$ to be specified shortly. By assuming (7) we omit the step that shows $\hat{\mathcal{I}}(\mathbf{a})$ constructed from certain kernel density estimators is uniformly convergent.

We must point out, however, that the proof of (7) itself may be technically involved depending on the distribution of the predictor $\mathbf{X}$. If the density $\mathbf{X}$ is bounded away from zero and is supported by a pre-compact set, then $\hat{\mathcal{I}}(\mathbf{a})$ can be constructed from kernel densities estimators, and the proof of (7) follows from Delecroix, Härdle and Hristache (2003, Section 2). In general, one may be able to prove (7) by working with an expanding sequence of subsets in the support of $\mathbf{X}$, as was done in Ai (1997). The demonstration of (7) under different scenarios exceeds the scope of the present paper, however, and will be left for further research.

Now let Θ be a compact set in $\mathbb{R}^p$, which serves as the space to be searched for the $\mathbf{a}$'s. Let

$$\Theta_1 = \{\mathbf{a} \in \Theta : a_p = 1\}, \quad \text{and} \quad \mathbf{a}_1 = \operatorname{argmax}\{\mathcal{I}(\mathbf{a}) : \mathbf{a} \in \Theta_1\},$$

where a_p is the pth component of $\mathbf{a}$. For $k = 2, \ldots, d$, define Θ_k and $\mathbf{a}_k$ recursively as

$$\Theta_k = \{\mathbf{a} \in \Theta_{k-1} : \mathbf{a} \perp \mathbf{a}_{k-1}\}, \quad \text{and} \quad \mathbf{a}_k = \operatorname{argmax}\{\mathcal{I}(\mathbf{a}) : \mathbf{a} \in \Theta_k\}.$$

Note that if $\mathbf{a}_1, \ldots, \mathbf{a}_d$ are finite, we can always choose Θ large enough to include all of them. Let

$$\begin{aligned}
&\hat{\mathbf{a}}_1 = \operatorname{argmax}\{\hat{\mathcal{I}}(\mathbf{a}) : \mathbf{a} \in \Theta_1\},\\
&\widehat{\Theta}_2 = \{\mathbf{a} \in \Theta_1 : \mathbf{a} \perp \hat{\mathbf{a}}_1\}, \quad \hat{\mathbf{a}}_2 = \operatorname{argmax}\{\hat{\mathcal{I}}(\mathbf{a}) : \mathbf{a} \in \widehat{\Theta}_2\}, \quad \text{and}\\
&\widehat{\Theta}_k = \{\mathbf{a} \in \widehat{\Theta}_{k-1} : \mathbf{a} \perp \hat{\mathbf{a}}_{k-1}\}, \quad \hat{\mathbf{a}}_k = \operatorname{argmax}\{\hat{\mathcal{I}}(\mathbf{a}) : \mathbf{a} \in \widehat{\Theta}_k\}, \quad k = 3, \ldots, d.
\end{aligned}$$

Thus we have population maximizers and sequential domains $\mathbf{a}_1, \ldots, \mathbf{a}_k$, $\Theta_1, \Theta_2, \ldots, \Theta_k$ of $\mathcal{I}(\mathbf{a})$ and their sample counterparts $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_k$, $\Theta_1, \widehat{\Theta}_2, \ldots, \widehat{\Theta}_k$ of $\hat{\mathcal{I}}(\mathbf{a})$. Let us stress that in the two sequences of domains, the first domains in both sequences are the same, but the latter domains in the two sequences differ, with $\Theta_2, \ldots, \Theta_d$ being fixed domains and $\widehat{\Theta}_2, \ldots, \widehat{\Theta}_d$ being random domains. We now state the results for consistency.

Proposition 10 *Suppose:*

1. Θ *is a compact set, and* $\mathcal{I}(\mathbf{a})$ *is continuous on* Θ.
2. $\sup\{|\hat{\mathcal{I}}(\mathbf{a}) - \mathcal{I}(\mathbf{a})| : \mathbf{a} \in \Theta_1\} \xrightarrow{p} 0$.
3. *The distribution* $p(Y|\mathbf{Z})$ *is directionally identifiable to the first order.*

Then $\hat{\mathbf{a}}_k \xrightarrow{p} \mathbf{a}_k$, *for* $k = 1, \ldots, d$.

To estimate $\mathcal{I}(\mathbf{a})$, we adopt the "leave-one-out" kernel density estimators used in Delecroix, Härdle and Hristache (2003). Let K denote a univariate kernel. Construct one- and two-dimensional density estimates as follows:

$$\hat{f}_i(u) = \frac{1}{(n-1)h_{11}} \sum_{j=1, j\neq i}^{n} K\left(\frac{u - u_j}{h_{11}}\right),$$

$$\hat{f}_i(u_1, u_2) = \frac{1}{(n-1)h_{21}h_{22}} \sum_{j=1, j\neq i}^{n} K\left(\frac{u_1 - u_{j1}}{h_{21}}\right) K\left(\frac{u_2 - u_{j2}}{h_{22}}\right),$$

where $h_{ij} = c_{ij} n^{-\delta_{ij}}$ with generic constants $c_{ij} > 0$ and $\delta_{ij} > 0$. In this paper the leave-one-out procedure will be used only for the theoretical purpose of ensuring that the argument of Delecroix, Härdle and Hristache (2003) is applicable, so that the uniform convergence of $\hat{\mathcal{I}}(\mathbf{a})$ is justified. Leaving out one observation in constructing a kernel density estimator is not very important in practice, and it is convenient and may be beneficial to use a full sample procedure, which is indeed how we implement our algorithm.

If the density of $\mathbf{X}$ is indeed bounded away from 0 and supported by a compact (or pre-compact) set, then we estimate $\mathcal{I}(\mathbf{a})$ by

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{\hat{f}_i(y_i, \mathbf{a}^T \mathbf{x}_i)}{\hat{f}_i(\mathbf{a}^T \mathbf{x}_i)} \right].$$

In general we use the truncated version

$$\frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{\hat{f}_i(y_i, \mathbf{a}^T \mathbf{x}_i)}{\hat{f}_i(\mathbf{a}^T \mathbf{x}_i)} I_{\{\mathbf{x}_i \in S\}} \right],$$

where S a large, compact, elliptical set in the support of $\mathbf{X}$ (or spherically shaped in the space of $\mathbf{Z}$). When we use the truncated estimator, treating S as fixed, we are estimating the CS $\mathcal{S}_{Y|\mathbf{X}_S}$, where $\mathbf{X}_S$ is $\mathbf{X}$ restricted to S, with distribution defined by

$$\Pr(\mathbf{X}_S \in A) = \Pr(\mathbf{X} \in A \cap S) / \Pr(\mathbf{X} \in S).$$

It can be shown that $\mathcal{S}_{Y|\mathbf{X}_S} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ for all S and $\mathcal{S}_{Y|\mathbf{X}_S} = \mathcal{S}_{Y|\mathbf{X}}$ for sufficiently large S. We state and prove this as Proposition 11 in the Appendix.

6 Examples

In this section, we present the results of a small simulation study and the analysis of a real data set, adding further results to the ones by Yin and Cook (2005).

6.1 Simulation Results

Here we present results based on simulated data from three models, each analysis being based on one data set.

Model 1 Consider the model

$$Y = \boldsymbol{\beta}^T \mathbf{X} + 0.5\epsilon,$$

where $\mathbf{X}$ is a random vector in $\mathbb{R}^5$ having a $N(0, I_5)$ distribution, $\boldsymbol{\beta}$ is the vector $(1, 2, 0, 0, 0)^T$. The sample size n is taken to be 100. This is a single index regression with a linear mean function.

The correlation coefficients between the true predictor and predictor found by our method was .9987. The permutation test gave a p-value of 0, indicating a clear first direction. In the orthogonal search for a second direction, the permutation test yielded a p-value of .981, correctly indicating that $d = 1$.

To see how the non-normality of the predictors might affect the result, we re-ran the above regression model with the predictors replaced by the independent random variables

$$X_1 \sim t_{(5)},\ X_2 \sim \chi^2_{(3)},\ X_3 \sim F(1, 4),\ X_4 \sim N(0, 1),\ X_5 \sim t_{(7)}.$$

In this case, the correlation coefficients between the true predictor and estimated one was .9987 (It's a coincidence that this is the same as the value observed with normal predictors). The permutation test p-values for the first and the second directions were 0 and .676, respectively. The non-normality of the predictors did not seem to have any notable impact on the results.

Model 2 Next, consider the model

$$Y = 0.5(\boldsymbol{\beta}^T \mathbf{X})^2 \epsilon,$$

where $\mathbf{X}$ is a 5-dimensional random vector with distribution $N(0, I_5)$, $\boldsymbol{\beta}$ is the vector $(2, 3, 0, 0, 0)^T$, and $\epsilon \sim N(0, 1)$. The sample size is taken to be $n = 200$. This is a single index regression with constant mean function but a variable variance function. The correlation coefficients between the true predictor and our estimated one was .9975. With a permutation test p-value of 0, the method again finds the true direction. The

permutation test from the orthogonal search for a second direction gave a p-value of .835. Again, we inferred the correct dimension and found a very good estimate of the CS.

Model 3 Now consider the "monkey saddle" model in Li (1992)

$$Y = X_1^3/3 - X_1 X_2^2,$$

where X_1 and X_2 are the first two components of a 5-dimensional random vector $\mathbf{X}$, which is distributed as $N(0, I_5)$. The sample size is taken to be $n = 200$. This is 2D regression, with both directions being in the mean function.

The multiple correlation coefficient between our first estimate and the two true predictors (X_1 and X_2) was .9966, with a permutation test p-value of 0. The first orthogonal search produced a second direction with a multiple correlation coefficient of .9981, and again gave a permutation test p-value of 0. The second orthogonal search resulted in a p-value of .531. Overall, we ended with the inference that $d = 2$, and a good estimate of the CS.

To see how the non-normality of the predictors affect our result, we replaced the predictors by the independent random variables

$$X_1 \sim N(0,1),\ X_2 \sim t_{(2)},\ X_3 \sim \chi^2_{(3)},\ X_4 \sim F(1,4),\ X_5 \sim \text{Gamma}(1,4).$$

The results were very similar to those from the normal-predictor case. We again inferred that $d = 2$ and ended with a very good estimate of the CS.

Model 4 Finally, consider a model used by Xia, et al. (2002, eq. (4.1))

$$Y = \boldsymbol{\beta}_1^T \mathbf{X}(\boldsymbol{\beta}_2^T \mathbf{X} + 1) + 0.5\epsilon,$$

where $\mathbf{X}$ is a 10-dimensional random vector with distribution $N(0, I_{10})$, $\boldsymbol{\beta}_1 = (1, 0, ..., 0)^T$, $\boldsymbol{\beta}_2 = (1, 1, 0, .., 0)^T$, and $\epsilon \sim N(0, 1)$. The sample size is taken to be $n = 200$. With $\mathbf{B}_0 = (\mathbf{b}_1, \mathbf{b}_2)$, where $\mathbf{b}_1 = (1, 0, ..., 0)^T$ and $\mathbf{b}_2 = (0, 1, 0, ..., 0)^T$, we used the criteria $m^2(\hat{\boldsymbol{\beta}}_1, \mathbf{B}_0)$ and $m^2(\hat{\boldsymbol{\beta}}_2, \mathbf{B}_0)$ from Section 2.1.1 of Xia, et al (2002) to summarize 100 replicates by our method. The resulting sample means of $m^2(\hat{\boldsymbol{\beta}}_1, \mathbf{B}_0)$ and $m^2(\hat{\boldsymbol{\beta}}_2, \mathbf{B}_0)$ were .0048 and .1049. Though our method is not particularly targeting the mean function, the results are similar to those in Figure 1(a) of Xia, et al (2002) where the mean function is the sole concern.

6.2 Automobile Collision Data

This is an automobile collision study data which consist of observations from 58 simulated side impact collision as described in Kallieris, Mattern and Härdle (1989). Of

interest is whether the accidents judged to result in a fatality, so the response is $Y = 1$ if fatal, $Y = 0$ if not fatal. The three predictor variables are the maximal acceleration measured on the subject's abdomen (X_1), age of the subject (X_2), and velocity of the automobile (X_3). The data also was studied by Härdle and Stoker (1989), and Yin and Cook (2005). Both of these studies assumed one dimensional structure. We continue the work of Yin and Cook (2005) by our successive search using permutation test. The first p-value is 0.000, while the second p-value is .797. Thus one-dimensional structure is confirmed.

7 Discussion

The fundamental message we would like to deliver is that the information extraction method for multiple index regressions, as a generalization of Yin and Cook (2005) for single index regressions, can be carried out one vector at a time if the predictor $\mathbf{X}$ has an elliptical distribution. It is the symmetry in $\mathbf{X}$ that enables us to deduce the fundamental conditional independence relation (3), which is the key to the Fisher consistency of the successive information extraction. Extracting information one vector at a time requires kernels of dimension no more than two in density estimation, thereby mitigating the local sparsity caused by high-dimensional kernel smoothing. This seems to further justify efforts to pre-process the predictors toward having an elliptical distribution by transformation or re-weighting (Cook and Nachtsheim, 1994).

Comparing with classical global methods such as OLS, SIR (Li, 1991), PHD (Li, 1992), and SAVE (Cook and Weisberg, 1991), our method for exhaustive estimation of the CS requires only that the distribution of $\mathbf{X}$ be elliptically contoured. Classical methods typically require additional conditions for exhaustive estimation. Another (potential) advantage is that our method is in essence the semi-nonparametric maximum likelihood estimator for dimension reduction carried out one direction at a time, and thus, following Ai (1997), it may be possible to show that it is semiparametrically efficient under some conditions. As a trade off, because it uses a two-dimensional kernel smoother, it likely requires larger sample sizes than the classical methods for its population and asymptotic advantages to take effect. In comparison, the classical methods only require slicing over a one-dimensional variable, and the slice width need not shrink as the sample size increases. This makes them unique among existing dimension reducing mechanisms.

Comparing with the more recent local methods such as MAVE (Xia et al., 2002) and structural adaptive estimation (SAE) (Hristache et al., 2001), the present method recovers the CS, whereas MAVE and SAE recover the central mean subspace, which is a subset of the CS. Furthermore, our method uses a two-dimensional kernel, whereas MAVE and SAE employ higher dimensional kernels, thus increasing the extent of local

sparsity in kernel smoothing.

Ai (1997) is a fundamental paper in the study of multiple index models. The present paper is a natural continuation in this direction, but there are important differences. The parameters $(\mathbf{a}_1, \ldots, \mathbf{a}_d)$ are estimated simultaneously in Ai's method, by solving a nonparametric score equation over a space of $p \times d$ matrices. In comparison, our method decomposes a multiple index model into a sequence of single index models, which, we believe, has important ramifications in the study of multiple index models.

First, the optimization procedure is simplified considerably: instead of solving a pd-dimensional score equation over a space of matrices in $\mathbb{R}^{p\times d}$, we need only carry out a sequence of d maximizations, each having dimension no more than p. Besides dimensionality, simultaneous estimation of $\mathbf{a}_1, \ldots, \mathbf{a}_d$ (by solving a score equation) is further complicated by the fact that, in the context considered in this paper, only the *column space* of $(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ is identifiable. That is, the (true) score for $\mathbf{a}_1, \ldots, \mathbf{a}_d$ is constant over the set

$$\{(\mathbf{a}_1, \ldots, \mathbf{a}_d)\mathbf{B} : \ \mathbf{B} \text{ is any nonsingular matrix in } \mathbb{R}^{d\times d}\}.$$

Consequently, the gradient matrix of the true score is singular and its rank is much less than its dimension, which is $pd \times pd$. standardize the $\mathbf{a}$'s does not help in this regard. The numerical behavior of the nonparametric score in Ai (1997), since it converges to the true score, would mimic the above singularity. Our sequential procedure avoids this problem.

Second, successive maximization provides us a natural ranking of the importance of the directions, which allowed us to develop a test procedure for determining d. To our knowledge such a procedure is not available for Ai's method.

Third, as mentioned before, the dimension of the kernel density estimator required by our method is at most two; whereas that required by Ai's method is $d + 1$. Without further asymptotic analysis we do not yet know how this difference will affect the two procedures asymptotically, but our intuition is that a lower-dimension kernel is better than a higher-dimensional kernel for finite samples.

Our approach also leads to two fundamental concepts that help streamline the current theoretical structure of dimension reduction: directional identifiability and M-sets. Directional identifiability, combined with information and the sufficiency of the CS, echoes the theoretical structure of classical statistical inference and allows us to adapt ideas therefrom, but at the same time takes into consideration the features special to the dimension reduction problem — that is, statistical inference depends only on the column space of the parameter matrix. The notion of M-sets, besides serving the purpose for which it was introduce, captures the essence of the existence of the CS, which lies at the foundation of sufficient dimension reduction.

Acknowledgment

The authors would like to thank valuable comments from an Editor and two referees, which greatly improve the paper.

Appendix: Technical details

The following equivalence, which can be found in Proposition 6.4 of Cook (1998), will be frequently used in various places: If $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3$ are random elements, then

$$\mathbf{U}_1 \perp\!\!\!\perp \mathbf{U}_2 | \mathbf{U}_3, \quad \mathbf{U}_1 \perp\!\!\!\perp \mathbf{U}_3 \text{ if and only if } \mathbf{U}_1 \perp\!\!\!\perp (\mathbf{U}_2, \mathbf{U}_3). \tag{8}$$

Proof of Proposition 3. Fix a $\mathbf{u}_3$, let

$$f(y, \mathbf{u}_1) = F_{Y|\mathbf{U}_1\mathbf{U}_3}(y|\mathbf{u}_1, \mathbf{u}_3) \text{ and } g(y, \mathbf{u}_2) = F_{Y|\mathbf{U}_2\mathbf{U}_3}(y|\mathbf{u}_2, \mathbf{u}_3).$$

The proposition will hold if, for any $(\mathbf{u}_1, \mathbf{u}_2), (\mathbf{u}_1', \mathbf{u}_2') \in \Omega_{12}(\mathbf{u}_3)$, we have

$$f(y, \mathbf{u}_1) = f(y, \mathbf{u}_1'), \quad g(y, \mathbf{u}_2) = g(y, \mathbf{u}_2'). \tag{9}$$

Because $\Omega_{12}(\mathbf{u}_3)$ is an M-set, there is a set of pairs in $\Omega_{12}(\mathbf{u}_3)$,

$$\left(\mathbf{u}_1^{(1)}, \mathbf{u}_2^{(1)}\right), \ldots, \left(\mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)}\right),$$

with $(\mathbf{u}_1^{(1)}, \mathbf{u}_2^{(1)})$ denoting $(\mathbf{u}_1, \mathbf{u}_2)$ and $(\mathbf{u}_1^{(m)}, \mathbf{u}_2^{(m)})$ denoting $(\mathbf{u}_1', \mathbf{u}_2')$, such that, for $i = 1, \ldots, m-1$, either $\mathbf{u}_1^{(i)} = \mathbf{u}_1^{(i+1)}$ or $\mathbf{u}_2^{(i)} = \mathbf{u}_2^{(i+1)}$. Hence either of the following is true

$$f(y, \mathbf{u}_1^{(i)}) = f(y, \mathbf{u}_1^{(i+1)}), \quad g(y, \mathbf{u}_2^{(i)}) = g(y, \mathbf{u}_2^{(i+1)}). \tag{10}$$

Equation (9) will be proved if *both* of the above equalities hold. Because $(\mathbf{u}_1^{(i)}, \mathbf{u}_2^{(i)}) \in \Omega_{12}(\mathbf{u}_3)$ and $(\mathbf{u}_1^{(i+1)}, \mathbf{u}_2^{(i+1)}) \in \Omega_{12}(\mathbf{u}_3)$, we have, by assumption 1,

$$f(y, \mathbf{u}_1^{(i)}) = g(y, \mathbf{u}_2^{(i)}) \text{ and } f(y, \mathbf{u}_1^{(i+1)}) = g(y, \mathbf{u}_2^{(i+1)}).$$

If the first equality in (10) holds, then the second equality in (10) also holds, because

$$g(y, \mathbf{u}_2^{(i)}) = f(y, \mathbf{u}_1^{(i)}) = f(y, \mathbf{u}_1^{(i+1)}) = g(y, \mathbf{u}_2^{(i+1)}).$$

Similarly, if the second equality in (10) holds then the first equality in (10) also holds. This completes the proof. □

PROOF OF PROPOSITION 4. Cases (a) and (b) follow from Definition 3. By the same definition, if $d = 2$ and $p(Y|\mathbf{Z})$ is first-order directionally identifiable then it must be directionally identifiable. For $d = 3$ we demonstrate by contradiction that if $p(Y|\mathbf{Z})$ is first order directionally identifiable then $p(Y|\mathbf{Z})$ must be second order directionally identifiable. The general case follows by induction using the same logic.

Assume then that $p(Y|\mathbf{Z})$ is first order directionally identifiable but not second order directionally identifiable. It follows from the definition of second-order directional identifiability and Corollary 1 that there are two distinct two-dimensional subspaces $\mathcal{S}_2 \subset \mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_2' \subset \mathbb{R}^p$ such that $Y \perp\!\!\!\perp P_{\mathcal{S}_2}\mathbf{Z}$ and $Y \perp\!\!\!\perp P_{\mathcal{S}_2'}\mathbf{Z}$. Consequently, we can construct two one-dimensional subspaces $\mathcal{S}_1 \subset \mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_1' \subset \mathbb{R}^p$ so that $Y \perp\!\!\!\perp P_{\mathcal{S}_1}\mathbf{Z}$ and $Y \perp\!\!\!\perp P_{\mathcal{S}_1'}\mathbf{Z}$. But this contradicts first order identifiability and consequently the conclusion follows. □

PROOF OF PROPOSITION 6. Let $\mathbf{a}$ be a vector in $\mathcal{S}^\perp \setminus (\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}})$. Note that this excludes the possibility that $\mathbf{a} = 0$.

First, suppose that $\mathbf{a}$ is not perpendicular to $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$, so that $P_1\mathbf{a} \neq 0$. Because $\mathbf{a}$ does not belong to $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$, it is linearly independent of $P_1\mathbf{a}$. Therefore, by the first-order directional identifiability, the ratio $p(Y|\mathbf{a}^T\mathbf{Z})/p(Y|(P_1\mathbf{a})^T\mathbf{Z})$ is a non-degenerate random variable. By Jensen's inequality,

$$\mathrm{E}\log\left[\frac{p(Y|\mathbf{a}^T\mathbf{Z})}{p(Y|(P_1\mathbf{a})^T\mathbf{Z})}\right] < \log\mathrm{E}\left[\frac{p(Y|\mathbf{a}^T\mathbf{Z})}{p(Y|(P_1\mathbf{a})^T\mathbf{Z})}\right].$$

Let $\mathbf{U} = \mathbf{a}^T\mathbf{Z}$, $\mathbf{V} = (P_1\mathbf{a})^T\mathbf{Z}$. Let μ denote the measure on the sample space Ω_Y of Y, with respect to which the density of Y is defined. Then the expectation in the expression on the right hand side is the integral

$$\int \frac{p(y|\mathbf{u})}{p(y|\mathbf{v})}\, p(\mathbf{u},\mathbf{v},y)\, d\mathbf{u}d\mathbf{v}d\mu(y) = \int p(y|\mathbf{u})\, p(\mathbf{u}|y,\mathbf{v})\, p(\mathbf{v})\, d\mathbf{u}d\mathbf{v}d\mu(y).$$

By (3), $p(\mathbf{u}|y,\mathbf{v}) = p(\mathbf{u}|\mathbf{v})$, and so the right hand side of the above equation becomes

$$\begin{aligned}\int p(y|\mathbf{u})\, p(\mathbf{u}|\mathbf{v})\, p(\mathbf{v})\, d\mathbf{u}d\mathbf{v}d\mu(y) &= \int p(y|\mathbf{u}) \left(\int p(\mathbf{u}|\mathbf{v})\, p(\mathbf{v})\, d\mathbf{v}\right) d\mathbf{u}d\mu(y) \\ &= \int p(y|\mathbf{u})\, p(\mathbf{u})\, d\mathbf{u}d\mu(y) = 1.\end{aligned}$$

Thus we have proved that $\mathcal{I}(\mathbf{b}) > \mathcal{I}(\mathbf{a})$ with $\mathbf{b} = P_1\mathbf{a} \neq 0$ and $\mathbf{b} \in \mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ for this case.

Next, suppose $\mathbf{a} \perp (\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}})$. Let $\mathbf{h}$ be any non-zero vector in $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$. Let $\mathbf{U} = \mathbf{a}^T\mathbf{Z}$ and $\mathbf{V} = \mathbf{h}^T\mathbf{Z}$. By (3), Y and $\mathbf{U}$ are independent. Hence

$$E\left[\frac{p(Y|\mathbf{U})}{p(Y|\mathbf{V})}\right] = E\left[\frac{p(Y)}{p(Y|\mathbf{V})}\right] = E\left[\frac{p(Y)p(\mathbf{V})}{p(Y,\mathbf{V})}\right] = 1. \tag{11}$$

Now apply Jensen's inequality and directional identifiability to complete the proof. □

Proof of Lemma 3.

1. Here, we note that $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$ or $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}^{\perp}$ implies that $\mathbf{a} \neq 0$. Because $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$ and $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}^{\perp}$, we have $Q_1\mathbf{a} \neq 0$ and $P_1\mathbf{a} \neq 0$. Since $p - d \geq 2$ and $Q_1 = Q$, there is a vector $\mathbf{v}$ in $\mathcal{S}(Q_1)$ such that $\mathbf{v} \perp Q_1\mathbf{a}$ and $\|\mathbf{v}\| = \|Q_1\mathbf{a}\|$. Because $\mathbf{v}$ and $Q_1\mathbf{a}$ are both in $\mathcal{S}(Q_1)$ they can be represented as $\mathbf{v} = \mathbf{W}\mathbf{c}_1$ and $Q\mathbf{a} = \mathbf{W}\mathbf{c}_2$, and by construction $\|\mathbf{c}_1\| = \|\mathbf{v}\| = \|Q_1\mathbf{a}\| = \|\mathbf{c}_2\|$. Let $\mathbf{A}$ be a $(p-d) \times (p-d)$ orthogonal matrix such that $\mathbf{c}_1 = \mathbf{A}\mathbf{c}_2$. This is possible because the orbit of $\mathbf{c}_2$, $\{\mathbf{A}\mathbf{c}_2 : \mathbf{A}$ is an orthogonal matrix$\}$, is the sphere $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = \|\mathbf{c}_2\|\}$. It follows that

$$\mathbf{W}\mathbf{c}_2 = \mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{W}\mathbf{c}_1, \quad \text{that is,} \quad \mathbf{v} = \mathbf{W}\mathbf{A}\mathbf{W}^T Q_1\mathbf{a} = \mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a}.$$

It is now easy to see that $\{\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a}, Q_1\mathbf{a}, P_1\mathbf{a}\}$ is an orthogonal set of vectors in $\mathbb{R}^{p-d}$. Hence they are linearly independent.

2. Proof is similar to the part (1). □

Proof of Proposition 7. Since none of the following arguments will be affected by a probability 0 set, we assume, without loss of generality, that $\Omega_{\mathbf{Z}}$ is itself open, so that it satisfies all the requirements for the set C in Lemma 1.

If $\mathbf{a} \in \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$ then $\mathbf{a} = P_1\mathbf{a}$ and (3) holds trivially. So for the rest of the proof assume $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$. Because $P_1 + Q_1$ is the projection onto $\mathcal{S}^{\perp}$ and $\mathbf{a} \in \mathcal{S}^{\perp}$, (3) is implied by

$$Y \perp\!\!\!\perp (Q_1\mathbf{a})^T\mathbf{Z} | (P_1\mathbf{a})^T\mathbf{Z}. \tag{12}$$

We now prove (12) in two cases.

Case I: $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}$ and $\mathbf{a} \notin \mathcal{S}^{\perp} \cap \mathcal{S}_{Y|\mathbf{Z}}^{\perp}$. Let $\mathbf{A}$ be as defined in Lemma 3, part 1. Then $\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a}, Q_1\mathbf{a}, P_1\mathbf{a}$ are linearly independent. By Lemma 1, for any $u \in \mathbb{R}$ such that $(P_1\mathbf{a})^T\mathbf{z} = u$ for some $\mathbf{z} \in \Omega_{\mathbf{Z}}$, the set

$$\left\{ \left(\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{z}, (Q_1\mathbf{a})^T\mathbf{z}\right) : (P_1\mathbf{a})^T\mathbf{z} = u, \mathbf{z} \in \Omega_{\mathbf{Z}} \right\}$$

is an M-set. Hence, by Proposition 3, (12) will hold if

$$p\left(Y | (\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}\right) = p\left(Y | (Q_1\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}\right).$$

In other words, we need to show that, for any measurable set C in the sample space Ω_Y of Y, we have

$$\mathrm{E}[I_C(Y) | (\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}] = \mathrm{E}[I_C(Y) | (Q_1\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}]. \tag{13}$$

We can replace $I_C(Y)$ by $\mathrm{E}(I_C(Y)|\mathbf{Z})$ in the above equality. However, because P is the projection onto the CS, we have $\mathrm{E}(I_C(Y)|\mathbf{Z}) = \mathrm{E}(I_C(Y)|P\mathbf{Z})$. In other words, equation (13) is equivalent to

$$\mathrm{E}[h(P\mathbf{Z})|(\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}] = \mathrm{E}[h(P\mathbf{Z})|(Q_1\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}],$$

where $h(P\mathbf{Z}) = \mathrm{E}(I_C(Y)|P\mathbf{Z})$. This equation will hold if

$$\left(h(P\mathbf{Z}), (\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}\right) \stackrel{d}{=} \left(h(P\mathbf{Z}), (Q_1\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}\right), \tag{14}$$

where, for two random vectors $\mathbf{V}_1$ and $\mathbf{V}_2$, $\mathbf{V}_1 \stackrel{d}{=} \mathbf{V}_2$ means that they have the same distribution. Let $\mathbf{B}$ be as defined in (4) and $\mathbf{Z}^* = \mathbf{B}\mathbf{Z}$. Because $\mathbf{Z}$ has a spherical distribution and $\mathbf{B}$ is an orthogonal matrix (Lemma 2), we have $\mathbf{Z}^* \stackrel{d}{=} \mathbf{Z}$. Hence

$$\begin{aligned}\left(h(P\mathbf{Z}), (\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (P_1\mathbf{a})^T\mathbf{Z}\right) &\stackrel{d}{=} \left(h(P\mathbf{Z}^*), (\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}^*, (P_1\mathbf{a})^T\mathbf{Z}^*\right) \\ &= \left(h(P\mathbf{B}\mathbf{Z}), (\mathbf{B}^{-1}\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}, (\mathbf{B}^{-1}P_1\mathbf{a})^T\mathbf{Z}\right). \end{aligned} \tag{15}$$

However, by Lemma 2, $P\mathbf{B} = P$, $\mathbf{B}^{-1}\mathbf{W}\mathbf{A}\mathbf{W}^T = Q_1$, and $\mathbf{B}^{-1}P_1 = P_1$. So the right hand sides of (14) and (15) are identical, which proves (14).

Case II: $\mathbf{a} \notin \mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ and $\mathbf{a} \in \mathcal{S}^\perp \cap \mathcal{S}^\perp_{Y|\mathbf{Z}}$. In this case $P_1\mathbf{a} = 0$, $Q_1\mathbf{a} = \mathbf{a}$, and (12) reduces to

$$Y \perp\!\!\!\perp \mathbf{a}^T\mathbf{Z}. \tag{16}$$

Let $\mathbf{A}$ be as defined in Lemma 3, part 2. Then $\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a}$ and $\mathbf{a}$ are linearly independent. Hence, by Lemma 1, the set

$$\left\{\left((\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{z}, \mathbf{a}^T\mathbf{z}\right) : \mathbf{z} \in \Omega_{\mathbf{Z}}\right\}$$

is an M-set. By Corollary 1 it suffices to show that

$$p(Y|(\mathbf{W}\mathbf{A}\mathbf{W}^T\mathbf{a})^T\mathbf{Z}) = p(Y|\mathbf{a}^T\mathbf{Z}).$$

The rest of the proof is parallel to that of Case I and will be omitted. □

Proof of Proposition 8. Let $\mathcal{S}$, P, Q, P_1, and Q_1 be as defined in Section 3.3, and recall that $Q = Q_1$. Let $\mathbf{\Sigma}_Y$ denote the conditional variance $\mathrm{Var}(\mathbf{Z}|Y)$. Let A be as defined in Lemma 4. We will first show (5). Recall that $\boldsymbol{\beta}$ is a $p \times d$ matrix whose columns form a basis of $\mathcal{S}_{Y|\mathbf{Z}}$. Let $\boldsymbol{\beta}_1$ denote a $p \times (p-d)$ matrix such that $\mathcal{S}(\boldsymbol{\beta}_1) = \mathcal{S}^\perp_{Y|\mathbf{Z}}$.

Suppose $\mathbf{Z}$ is normal. Then, by our convention $\mathbf{Z}$ is distributed as $N(0, I_p)$. We have $P\mathbf{Z} \perp\!\!\!\perp Q\mathbf{Z}$. Moreover, we know that $Y \perp\!\!\!\perp \mathbf{Z}|P\mathbf{Z}$, and hence $Y \perp\!\!\!\perp Q\mathbf{Z}|P\mathbf{Z}$. Then, by equivalence (8), relation (5) holds.

If $\mathbf{Z}|Y$ is normal, then by parts 1 and 3 of Lemma 4, $\text{span}\{I_p - \mathbf{\Sigma}_y : y \in A\} = \mathcal{S}_{Y|\mathbf{Z}}$, which implies $Q(I_p - \mathbf{\Sigma}_y) = 0$ for $y \in A$. Multiply both sides of this equation from the right by P and Q, respectively, to obtain

$$Q\mathbf{\Sigma}_y P = QP = 0 \;\text{ and }\; Q\mathbf{\Sigma}_y Q = Q \;\text{ for } y \in A.$$

The first equality implies $P\mathbf{Z} \perp\!\!\!\perp Q\mathbf{Z}|\{Y = y\}$ for all $y \in A$ which, because $P(Y \in A) = 1$, implies that $P\mathbf{Z} \perp\!\!\!\perp Q\mathbf{Z}|Y$. The second equality implies $\text{Var}(Q\mathbf{Z}|Y = y) = Q$ for $y \in A$. Meanwhile, note that

$$I_p = \text{E}(\mathbf{\Sigma}_Y) + \text{Var}(\text{E}(\mathbf{Z}|Y)), \;\text{ or equivalently, }\; \text{E}(I_p - \mathbf{\Sigma}_Y) = \text{Var}(\text{E}(\mathbf{Z}|Y)).$$

Multiply the second equation from both sides by Q to obtain $\text{Var}(\text{E}(Q\mathbf{Z}|Y)) = 0$. Hence $\text{E}(Q\mathbf{Z}|Y = y) = 0$ on a set $A_1 \in \Omega_Y$ with $P(Y \in A_1) = 1$. Thus we have shown that $Q\mathbf{Z}|\{Y = y\}$ is distributed as $N(0, Q)$ for $y \in A \cap A_1$. Because the conditional distribution of $Q\mathbf{Z}|\{Y = y\}$ is constant on $y \in A \cap A_1$, which has probability 1, we see that $Q\mathbf{Z} \perp\!\!\!\perp Y$. This, combined with $P\mathbf{Z} \perp\!\!\!\perp Q\mathbf{Z}|Y$, implies (5) by the equivalence relation (8).

Now let $\mathbf{a}$ be a vector in $\mathcal{S}^\perp$. Because $\mathcal{S}^\perp \cap \mathcal{S}_{Y|\mathbf{Z}}$ is a subspace of $\mathcal{S}_{Y|\mathbf{Z}}$, $(P_1\mathbf{a})^T\mathbf{Z}$ is measurable with respect to $P\mathbf{Z}$. Because $Q_1 = Q$, $(Q_1\mathbf{a})^T\mathbf{Z}$ is measurable with respect to $Q\mathbf{Z}$. Consequently $(Q_1\mathbf{a})^T\mathbf{Z} \perp\!\!\!\perp (Y, (P_1\mathbf{a})^T\mathbf{Z})$, from which it follows that

$$Y \perp\!\!\!\perp (Q_1\mathbf{a})^T\mathbf{Z}|(P_1\mathbf{a})^T\mathbf{Z}.$$

Because, conditioning on $(P_1\mathbf{a})^T\mathbf{Z}$, $(P_1\mathbf{a})^T\mathbf{Z}$ is a constant, the above implies

$$Y \perp\!\!\!\perp ((P_1\mathbf{a})^T\mathbf{Z}, (Q_1\mathbf{a})^T\mathbf{Z})|(P_1\mathbf{a})^T\mathbf{Z}. \tag{17}$$

However, because $\mathbf{a} \in \mathcal{S}^\perp$ and $P_1 + Q_1$ is the orthogonal projection onto $\mathcal{S}^\perp$, we have $\mathbf{a}^T\mathbf{Z} = (P_1\mathbf{a})^T\mathbf{Z} + (Q_1\mathbf{a})^T\mathbf{Z}$. Thus (3) follows from (17). □

Proof of Proposition 9. Write $\text{Var}(\mathbf{Z}|Y)$ as $\mathbf{\Sigma}_Y$ and $E(\mathbf{Z}|Y)$ as $\boldsymbol{\mu}_Y$. By (8),

$$\mathbf{A}_{p-k}^T\mathbf{Z} \perp\!\!\!\perp (Y, \mathbf{A}_k^T\mathbf{Z}) \tag{18}$$

implies (6). We now show that (6) implies (18). Suppose $\mathbf{Z}$ is normal. Then it is distributed as $N(0, I_p)$. Because $\mathbf{A}_{p-k}^T\mathbf{A}_k = 0$, we have $\mathbf{A}_{p-k}^T\mathbf{Z} \perp\!\!\!\perp \mathbf{A}_k^T\mathbf{Z}$. Hence (18) follows from (8). Now suppose $\mathbf{Z}|Y$ is normal. By (6) we have $\mathcal{S}_{Y|\mathbf{Z}} \subseteq \mathcal{S}(\mathbf{A}_k)$. Hence, by Lemma 4, part 3, $\mathcal{S}(I - \mathbf{\Sigma}_Y) \subseteq \mathcal{S}(\mathbf{A}_k)$, which implies $\mathbf{A}_{p-k}^T(I - \mathbf{\Sigma}_Y) = 0$. It follows that

$$\mathbf{A}_{p-k}^T\mathbf{\Sigma}_Y\mathbf{A}_{p-k} = \mathbf{A}_{p-k}^T\mathbf{A}_{p-k} = I_{p-k} \;\text{ and }\; \mathbf{A}_{p-k}^T\mathbf{\Sigma}_Y\mathbf{A}_k = \mathbf{A}_{p-k}^T\mathbf{A}_k = 0. \tag{19}$$

In the meantime, by Lemma 4, part 1, $\boldsymbol{\mu}_Y$ belongs to $\mathcal{S}(I - \boldsymbol{\Sigma}_Y)$, a subspace of $\mathcal{S}(\mathbf{A}_k)$. Hence

$$\mathbf{A}_{p-k}^T \boldsymbol{\mu}_Y = 0. \tag{20}$$

From the second equality of (19), we see that $\mathbf{A}_{p-k}^T \mathbf{Z} \perp\!\!\!\perp \mathbf{A}_k^T \mathbf{Z} | Y$. By (20) and the first equality of (19), $\mathbf{A}_{p-k}^T \mathbf{Z} | Y$ is distributed as $N(0, I_{p-k})$ and is therefore independent of Y. Now (18) follows from (8).

Next, suppose $p - d \geq 1$ and $p(Y|\mathbf{Z})$ is directionally identifiable. By the first part of this proposition it is easy to see that (6) implies

$$Y \perp\!\!\!\perp \mathbf{A}_{p-k}^T \mathbf{Z}. \tag{21}$$

We now show that (21) implies (6). It suffices to show that $\mathbf{a}_\ell$ belongs to $\mathcal{S}_{Y|\mathbf{Z}}^\perp$ for all $\ell = k+1, \ldots, p$. By construction, $\mathbf{a}_\ell$ belongs to either $\mathcal{S}_{Y|\mathbf{Z}}$ or its orthogonal complement $\mathcal{S}_{Y|\mathbf{Z}}^\perp$. Suppose $\mathbf{a}_\ell \in \mathcal{S}_{Y|\mathbf{Z}}$. Let $\mathbf{b}$ be any nonzero vector in $\mathcal{S}_{Y|\mathbf{Z}}^\perp$. We know, by Proposition 8, $Y \perp\!\!\!\perp \mathbf{b}^T \mathbf{Z}$. Hence

$$p(Y|\mathbf{a}_\ell^T \mathbf{Z}) = p(Y) = p(Y|\mathbf{b}^T \mathbf{Z}),$$

contradicting to the assumption of directional identifiability. □

Proof of Proposition 10. First, we show that $\hat{\mathbf{a}}_1 \xrightarrow{p} \mathbf{a}_1$. Let $\epsilon > 0$. By directional identifiability of $p(Y|\mathbf{Z})$, $\mathcal{I}(\mathbf{a})$ has a unique maximizer in Θ_1. Moreover, because Θ_1 is compact, there is a number $\delta > 0$ such that

$$\sup\{\mathcal{I}(\mathbf{a}) : \|\mathbf{a} - \mathbf{a}_1\| > \epsilon, \mathbf{a} \in \Theta_1\} < \mathcal{I}(\mathbf{a}_1) - \delta. \tag{22}$$

By assumption 2, we have, with probability tending to 1,

$$|\hat{\mathcal{I}}(\mathbf{a}_1) - \mathcal{I}(\mathbf{a}_1)| < \delta/4 \quad \text{and} \quad |\hat{\mathcal{I}}(\hat{\mathbf{a}}_1) - \mathcal{I}(\hat{\mathbf{a}}_1)| < \delta/4.$$

By construction, $\hat{\mathcal{I}}(\hat{\mathbf{a}}_1) \geq \hat{\mathcal{I}}(\mathbf{a}_1)$, which, combined with the first of the above inequalities, implies that, with probability tending to 1, $\hat{\mathcal{I}}(\hat{\mathbf{a}}_1) > \mathcal{I}(\mathbf{a}_1) - \delta/4$. Using the second of the inequalities we see that, with probability tending to 1,

$$\mathcal{I}(\hat{\mathbf{a}}_1) > \mathcal{I}(\mathbf{a}_1) - \delta/2.$$

By (22),

$$P\left(\|\hat{\mathbf{a}}_1 - \mathbf{a}_1\| > \epsilon\right) \leq P\left(\mathcal{I}(\hat{\mathbf{a}}_1) < \mathcal{I}(\mathbf{a}_1) - \delta\right).$$

Recall that if A_n and B_n be two sequences of events with $P(A_n) \to 1$, then $\limsup P(B_n) = \limsup P(B_n \cap A_n)$. Hence $\limsup P\left(\|\hat{\mathbf{a}}_1 - \mathbf{a}_1\| > \epsilon\right)$ is no more than

$$\begin{aligned}
&\limsup_{n\to\infty} P\left(\mathcal{I}(\hat{\mathbf{a}}_1) < \mathcal{I}(\mathbf{a}_1) - \delta\right) \\
&= \limsup_{n\to\infty} P\left(\mathcal{I}(\hat{\mathbf{a}}_1) < \mathcal{I}(\mathbf{a}_1) - \delta, \mathcal{I}(\hat{\mathbf{a}}_1) > \mathcal{I}(\mathbf{a}_1) - \delta/2\right) = 0.
\end{aligned}$$

Thus we have proved $\hat{\mathbf{a}}_1 \xrightarrow{p} \mathbf{a}_1$.

Next, we consider the proof of $\hat{\mathbf{a}}_k \xrightarrow{p} \mathbf{a}_k$ for $k = 2, \ldots, d$. For simplicity we will only prove $\hat{\mathbf{a}}_2 \xrightarrow{p} \mathbf{a}_2$, because this will reveal the full logic pattern needed for a general proof. Fix an $\epsilon > 0$. We need to show that

$$\limsup_{n\to\infty} P\left(\|\hat{\mathbf{a}}_2 - \mathbf{a}_2\| > \epsilon\right) = 0.$$

For a vector $\mathbf{a} \in \Theta_1$, let $\Theta_2(\mathbf{a})$ be the set $\{\mathbf{a}' \in \Theta_1 : \mathbf{a}' \perp \mathbf{a}\}$. In this notation $\widehat{\Theta}_2 = \Theta_2(\hat{\mathbf{a}}_1)$ and $\Theta_2 = \Theta_2(\mathbf{a}_1)$. By the continuity of inner product, the set $\{\mathbf{a}' \in \mathbb{R}^p : \mathbf{a}' \perp \mathbf{a}_1\}$ is closed. Consequently Θ_2 is compact. By directional identifiability $\mathcal{I}$ has a unique maximizer over Θ_2, which, combined with the compactness of Θ_2, implies that for some $\delta > 0$,

$$\sup\{\mathcal{I}(\mathbf{a}') : \|\mathbf{a}' - \mathbf{a}_2\| > \epsilon, \mathbf{a}' \in \Theta_2\} < \mathcal{I}(\mathbf{a}_2) - \delta.$$

Because $\mathcal{I}$ is continuous on Θ_1 and Θ_1 is compact, it is uniformly continuous on Θ_1. Hence, for all $\mathbf{a}$ sufficiently close to $\mathbf{a}_1$, say $\|\mathbf{a} - \mathbf{a}_1\| < \epsilon_1$, we have

$$\sup\{\mathcal{I}(\mathbf{a}') : \|\mathbf{a}' - \mathbf{a}_2\| > \epsilon, \mathbf{a}' \in \Theta_2(\mathbf{a})\} < \mathcal{I}(\mathbf{a}_2) - \delta/2.$$

Now, we know, with probability tending to 1, $\|\hat{\mathbf{a}}_1 - \mathbf{a}_1\| < \epsilon_1$. Hence

$$\begin{aligned}
\limsup_{n\to\infty} P\left(\|\hat{\mathbf{a}}_2 - \mathbf{a}_2\| > \epsilon\right) &= \limsup_{n\to\infty} P\left(\|\hat{\mathbf{a}}_2 - \mathbf{a}_2\| > \epsilon, \|\hat{\mathbf{a}}_1 - \mathbf{a}_1\| < \epsilon_1\right) \\
&\le \limsup_{n\to\infty} P\left(\mathcal{I}(\hat{\mathbf{a}}_2) < \mathcal{I}(\mathbf{a}_2) - \delta/2\right). \qquad (23)
\end{aligned}$$

Let $\mathbf{a}^*(\mathbf{a})$ be the closest point in the set $\Theta_2(\mathbf{a})$ to $\mathbf{a}_2$. Then $\lim_{\mathbf{a}\to\mathbf{a}_1} \|\mathbf{a}^*(\mathbf{a}) - \mathbf{a}_2\| = 0$. Thus, with probability tending to 1,

$$\|\mathcal{I}(\mathbf{a}^*(\hat{\mathbf{a}}_1)) - \mathcal{I}(\mathbf{a}_2)\| < \delta/8.$$

By assumption 2, we also know that

$$\|\hat{\mathcal{I}}(\mathbf{a}^*(\hat{\mathbf{a}}_1)) - \mathcal{I}(\mathbf{a}^*(\hat{\mathbf{a}}_1))\| < \delta/8 \quad \text{and} \quad \|\hat{\mathcal{I}}(\hat{\mathbf{a}}_2) - \mathcal{I}(\hat{\mathbf{a}}_2)\| < \delta/8.$$

However, by construction, $\hat{\mathcal{I}}(\hat{\mathbf{a}}_2) \geq \hat{\mathcal{I}}(\mathbf{a}^*(\hat{\mathbf{a}}_1))$. Hence, with probability tending to 1,

$$\mathcal{I}(\hat{\mathbf{a}}_2) > \hat{\mathcal{I}}(\hat{\mathbf{a}}_2) - \delta/8 > \hat{\mathcal{I}}(\mathbf{a}^*(\hat{\mathbf{a}}_1)) - \delta/8 > \mathcal{I}(\mathbf{a}^*(\hat{\mathbf{a}}_1)) - 2\delta/8 > \mathcal{I}(\mathbf{a}_2) - 3\delta/8. \quad (24)$$

Combine (23) and (24) to obtain

$$\begin{aligned}\limsup_{n\to\infty} P\left(\|\hat{\mathbf{a}}_2 - \mathbf{a}_2\| > \epsilon\right) &\leq \limsup_{n\to\infty} P\left(\mathcal{I}(\hat{\mathbf{a}}_2) < \mathcal{I}(\mathbf{a}_2) - \delta/2\right) \\ &\leq \limsup_{n\to\infty} P\left(\mathcal{I}(\hat{\mathbf{a}}_2) < \mathcal{I}(\mathbf{a}_2) - \delta/2, \mathcal{I}(\hat{\mathbf{a}}_2) > \mathcal{I}(\mathbf{a}_2) - 3\delta/8\right) \\ &= 0,\end{aligned}$$

which completes the proof. □

Proposition 11 *The following assertions hold:*

1. *If S is any (Borel) set in $\mathbb{R}^p$, then $\mathcal{S}_{Y|\mathbf{X}_S} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.*
2. *There is a compact set S of $\mathbb{R}^p$, which can be taken as an ellipsoid, such that $\mathcal{S}_{Y|\mathbf{X}_S} = \mathcal{S}_{Y|\mathbf{X}}$, and for any (Borel) set S' containing S, $\mathcal{S}_{Y|\mathbf{X}_{S'}} = \mathcal{S}_{Y|\mathbf{X}}$.*

Proof. 1. We will prove the stronger result: if $S \subseteq S'$ are two Borel sets in $\mathbb{R}^p$, then $\mathcal{S}_{Y|\mathbf{X}_S} \subseteq \mathcal{S}_{Y|\mathbf{X}_{S'}}$, which implies assertion 1 if we take $S' = \Omega_{\mathbf{X}}$. We need this general result for proving assertion 2. Let W be the indicator that takes value 1 if $\mathbf{X}_{S'} \in S$ and 0 if $\mathbf{X}_{S'} \in S' \setminus S$. Because W is a function of $\mathbf{X}_{S'}$, we have $Y \perp\!\!\!\perp W|\mathbf{X}_{S'}$. By Propositions 3.2 and 3.3 of Chiaromonte, Cook and Li (2002), we have $\mathcal{S}_{Y|\mathbf{X}_S} \subseteq \mathcal{S}_{Y|\mathbf{X}_{S'}}$.

2. Any random vector is tight. That is, for any $\epsilon > 0$, there is a compact set K such that $\Pr(\mathbf{X} \notin K) < \epsilon$. So there is a sequence of compact sets $S_1, S_2, \ldots$, with $S_1 \subseteq S_2 \subseteq \cdots$, such that $\Pr(\mathbf{X} \notin S_k) \to 0$ as $k \to \infty$. Let $\phi(\mathbf{t}, \tau) = E(e^{i\mathbf{t}^T\mathbf{X} + i\tau Y})$ and $\phi_k(\mathbf{t}, \tau) = E(e^{i\mathbf{t}^T\mathbf{X}_{S_k} + i\tau Y})$ be the characteristic functions of $(\mathbf{X}, Y)$ and $(\mathbf{X}_{S_k}, Y)$, respectively. Then

$$\begin{aligned}|\phi_k(\mathbf{t}, \tau) - \phi(\mathbf{t}, \tau)| &\leq E|e^{i\mathbf{t}^T\mathbf{X}_{S_k} + i\tau Y} - e^{i\mathbf{t}^T\mathbf{X} + i\tau Y}| \\ &\leq \int |e^{i\mathbf{t}^T\mathbf{X}_{S_k}} - e^{i\mathbf{t}^T\mathbf{X}}| dP \leq \int_{\Omega_{\mathbf{X}} \setminus S_k} |1 - e^{i\mathbf{t}^T\mathbf{X}}| dP \leq 2\Pr(\mathbf{X} \notin S_k) \to 0.\end{aligned}$$

Thus $(\mathbf{X}_{S_k}, Y)$ converges in distribution to $(\mathbf{X}, Y)$, and hence the conditional distribution of $Y|\mathbf{X}_{S_k}$ converges (almost surely) to that of $Y|\mathbf{X}$. This implies that $\mathcal{S}_{Y|\mathbf{X}_{S_k}}$ converges to $\mathcal{S}_{Y|\mathbf{X}}$. That is, if P_k and P be the projections on to $\mathcal{S}_{Y|\mathbf{X}_{S_k}}$ and $\mathcal{S}_{Y|\mathbf{X}}$, respectively, then $\|P_k - P\| \to 0$, where $\|\cdot\|$ is, say, the Frobenius matrix norm.

We now show that, for sufficiently large k, $P_k = P$. If not, then, for any k, there is a $k' \geq k$ such that $P_{k'} \neq P$. However, this implies, by assertion 1, that $\mathcal{S}_{Y|\mathbf{X}_{S_k}}$ is a

subspace of $\mathcal{S}_{Y|\mathbf{X}}$ for each k. In other words the rank of P_k must be smaller than that of P for each k, contradicting to the fact that $\|P_k - P\| \to 0$.

Now take S in assertion 2 to be any S_k for which $\mathcal{S}_{Y|\mathbf{X}_{S_k}} = \mathcal{S}_{Y|\mathbf{X}}$. Then, by assertion 1, $\mathcal{S}_{Y|\mathbf{X}_{S'}} = \mathcal{S}_{Y|\mathbf{X}}$ whenever S' contains S. The set S can be taken as an ellipsoid because any compact set in $\mathbb{R}^p$ is bounded, and is therefore contained in an ellipsoid. □

References

Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica*, 65, 933–963.

Chiaromonte, F., and Cook, R. D. (2002). Sufficient dimension reduction and graphics in regression. *The Ann. Inst. Statist. Math.* **54**, 768–795.

Chiaromonte, F., Cook, R. D. and Li. B. (2002). Dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, **30**, 475–497.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89, 177–190.

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

Cook, R. D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *Journal of American Statistical Association,* **94**, 1187–1200.

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

Cook, R. D. and Li, B. (2004). Determining the dimension of Iterative Hessian Transformation. *The Annals of Statistics*, 32, 2501–2531.

Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89, 592–599.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100, 410–428.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86, 328–332.

Cook, R. D. and Weisberg, S. (1999). Graphs in statistical analyses: Is the medium the message? *The American Statistician*, 53, 29–37.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, Vol. 43, No. 2, 147–199.

Cox, D.R. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, **86**, 213–226.

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

Härdle, W. and Stoker, T. M. (1989). Investigating Smooth Multiple Regression by the Method of Average Derivatives. *Journal of the American Statistical Association*, 84, 986–995.

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

Hristache, M., Juditsky, and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, **29**, 595–623.

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

Kalliers, D. Mattern, R. and Härdle, W. (1989). Verhalten des EUROSID beim 90 grad seitenaufprall im vergleich zu PMTO sowie US-SID, HYBRID II und APROD. in *Forschungsvereinigung Automobiltechnik (FAT) Schriftenreihe,* Frankfurt am Main.

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

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

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

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89, 141–148.

Xia, Y., Tong, H., Li, W. K. and Zhu, L-X. (2002). An adaptive estimation of dimension reduction space, *Journal of the Royal Statistical Society, Ser. B*, **64**, 363-410.

Xia, Y. (2007). A Constructive Approach to the Estimation of Dimension Reduction Directions. http://arxiv.org/abs/math/0701761

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

Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regression. *Biometrika*, 92, 371–384.

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