A general sufficient dimension reduction approach via Hellinger integral of order two

Qin Wang* Xiangrong Yin† Frank Critchley‡

Abstract

Sufficient dimension reduction provides a useful tool to study the dependence between a response and a multidimensional predictor. In this paper, a new formulation is proposed based on the Hellinger integral of order two – and so jointly local in the response and predictor – together with an efficient estimation algorithm. Our approach has a number of strengths. It requires minimal (essentially, just existence) assumptions. Relative to existing methods, it is computationally efficient while overall performance is broadly comparable, allowing larger problems to be tackled, more general, multidimensional response being allowed. A sparse version enables variable selection. Finally, it unifies three existing methods, each being shown to be equivalent to adopting suitably weighted forms of the Hellinger integral of order two.

Key Words: Central Subspace; Hellinger Integral; Sufficient Dimension Reduction.

*Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, USA. E-mail: qwang3@vcu.edu

†Department of Statistics, 204 Statistics Building, The University of Georgia, USA. E-mail: xryin@stat.uga.edu

‡Department of Mathematics and Statistics, The Open University, UK. E-mail: f.critchley@open.ac.uk

1 Introduction

In simple regression a 2-dimensional plot of the response Y versus the predictor 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. Sufficient dimension reduction (SDR) seeks low-dimensional analogues of this fully informative plot for a general $p \times 1$ predictor vector X, without pre-specifying a model for any of $Y|X$ or $X|Y$. That is, reducing the dimension of the predictors without loss of information on the conditional distribution of $Y|X$. Such a reduced predictor space is called a dimension reduction subspace for the regression of Y on X. We assume throughout that the intersection of all such spaces is itself a dimension reduction subspace, as holds under very mild conditions (Cook 1998a; Yin, Li and Cook 2008). This intersection, called the central subspace $\mathcal{S}_{Y|X}$ for the regression of Y on X (Cook 1994, 1996), becomes the natural focus of inferential interest. Its dimension $d_{Y|X}$ is called the structural dimension.

Since the first moment-based methods, *sliced inverse regression* (SIR; Li 1991) and *sliced average variance estimation* (SAVE; Cook and Weisberg 1991), were introduced, many others have been proposed. These methods can be categorized into three groups, according to which distribution is focused on: the inverse regression approach, the forward regression approach and the joint approach. Inverse regression methods focus on the inverse conditional distribution of $X|Y$. Alongside SIR and SAVE, *principal Hessian directions* (PHD; Li 1992; Cook 1998b), *parametric inverse regression* (Bura and Cook 2001), *the k^{th} moment estimation* (Yin and Cook 2002), *sliced average third-moment estimation* (Yin and Cook 2003), *inverse regression* (Cook and Ni 2005) and *contour regression* (Li, Zha and Chiaromonte 2005) are well-known approaches in this category among others. They are computationally inexpensive, but require either or both of the key linearity and constant covariance conditions (Cook 1998a). An exhaustiveness condition (recovery of the whole central subspace) is also required by some

of these methods. *Average derivative estimation* (Härdle and Stoker 1989; Samarov 1993), *the structure adaptive method* (Hristache, Juditsky, Polzehl and Spokoiny 2001), *minimum average variance estimation* (MAVE; Xia, Tong, Li and Zhu 2002), *sliced regression* (SR; Wang and Xia 2008), *ensemble estimation* (Yin and Li 2011) are examples of forward regression methods, where the conditional distribution of $Y|X$ is the object of inference. These methods do not require any strong probabilistic assumptions, but the computational burden increases dramatically with either sample size or the number of predictors, due to the use of nonparametric estimation. The third class – the joint approach – includes *Kullback-Leibler distance* (Yin and Cook 2005; Yin, Li and Cook 2008), *Fourier estimation* (Zhu and Zeng 2006) and *integral estimation* (Zeng and Zhu 2010), which may be flexibly regarded as either inverse or forward methods.

In this paper, we introduce a new approach that targets on the central subspace by exploiting a characterization of dimension reduction subspaces in terms of the Hellinger integral of order two. The assumptions needed are very mild: (a) $\mathcal{S}_{Y|X}$ exists, so we have a well-defined problem to solve and (b) a finiteness condition, so that the Hellinger integral is always defined, as holds without essential loss. Accordingly, our approach is more flexible than many others, multidimensional Y being allowed. Incorporating appropriate weights, it also unifies three existing methods, including SR.

The rest of the article is organized as follows. Section 2 introduces the new approach, including motivation and connection with dimension reduction. Section 3 covers its implementation, a k-nearest neighborhood (KNN) approximation of the Hellinger integral of order two. A sparse version is also described, enabling variable selection. Examples on both real and simulated data are given in Section 4. Final comments and some further developments are given in Section 5. Additional proofs and related materials can be found in the Appendix. Matlab codes for our algorithms are available upon request.

2 The Hellinger integral of order two

2.1 Notation and definition

We assume throughout that the response variable Y and the $p\times 1$ predictor vector X have a joint distribution $F_{(Y,X)}$, and that the data $\{(y_i,\ x_i), i = 1, \ldots, n\}$, are independent observations from it. Refer $p(w_1, w_2)$, $p(w_1|w_2)$ and $p(w_2)$ to the joint, conditional and marginal distributions of (W_1, W_2), $W_1|W_2$ and W_2 respectively.

The notation $W_1 \perp\!\!\!\perp W_2|W_3$ means that the random vectors W_1 and W_2 are independent given any value of the random vector W_3. Subspaces are usually denoted by $\mathcal{S}$. $P_{\mathcal{S}}$ denotes the orthogonal projection operator onto $\mathcal{S}$ in the usual inner product. For any x, $x_{\mathcal{S}}$ denotes its projection $P_{\mathcal{S}}x$. $\mathcal{S}(B)$ denotes the subspace of $\mathbb{R}^s$ spanned by the columns of the $s \times t$ matrix B. For B_i of order $s \times t_i$ $(i = 1,\ 2)$, $(B_1,\ B_2)$ denotes the matrix of order $s \times (t_1 + t_2)$ formed in the obvious way. Finally, $\mathcal{A} \subset \mathcal{B}$ means that $\mathcal{A}$ is a proper subset of $\mathcal{B}$, and $\mathcal{A} \subseteq \mathcal{B}$ indicates that $\mathcal{A}$ is a subset of $\mathcal{B}$, either $\mathcal{A} \subset \mathcal{B}$ or $\mathcal{A} = \mathcal{B}$.

Throughout, u, u_1, u_2, ... denote fixed matrices with p rows. The Hellinger integral H of order two is defined by $H(u) := \mathbb{E}\left\{R(Y; u^T X)\right\}$, where $R(y; u^T x)$ is the so-called *dependence ratio* $R(y; u^T x) = \frac{p(y,u^T x)}{p(y)p(u^T x)} = \frac{p(y|u^T x)}{p(y)} = \frac{p(u^T x|y)}{p(u^T x)}$, and the expectation is over the joint distribution, a fact which can be emphasized by writing $H(u)$ more fully as $H(u; F_{(Y,X)})$.

We assume $F_{(Y,X)}$ is such that $H(u)$ is finite for all u, so that Hellinger integrals are always defined. This finiteness condition is required without essential loss. It holds whenever Y takes each of a finite number of values with positive probability, a circumstance from which any sample situation is indistinguishable. Again, we know of no theoretical departures from it which are likely to occur in statistical practice, if only because of errors of observation. For example, if (Y, X) is bivariate normal with correlation ρ, $H(1) = \left(1 - \rho^2\right)^{-1}$ becomes infinite in either singular limit $\rho \to \pm 1$ but, then,

Y is a deterministic function of X.

2.2 Properties

We now study the properties of Hellinger integral of order two. Immediately, we can see that R and/or H can be viewed as forward regression, inverse regression and general correlation of Y on $u^T X$, while the invariance $\mathcal{S}_{Y^*|X} = \mathcal{S}_{Y|X}$ of the central subspace under any 1-1 transformation $Y \to Y^*$ of the response (Cook, 1998a) is mirrored locally in $R(y^*; u^T x) = R(y; u^T x)$ and hence, globally in $H(u; F_{(Y,X)}) = H(u; F_{(Y^*,X)})$. Furthermore, the relation $\mathcal{S}_{Y|Z} = A^{-1}\mathcal{S}_{Y|X}$ between central subspaces before and after nonsingular affine transformation $X \to Z := A^T X + b$ (Cook, 1998a) is mirrored locally in $R(y; u^T x) = R(y; (A^{-1}u)^T z)$ and hence, globally in $H(u; F_{(Y,X)}) = H(A^{-1}u; F_{(Y,Z)})$. This implies that one can freely use the scale of predictors.

Our first result establishes that $H(u)$ depends on u only via the subspace spanned by its columns.

Proposition 1 *If* $\mathrm{Span}(u_1) = \mathrm{Span}(u_2)$*, then* $H(u_1) = H(u_2)$*.*

Our primary interest is in subspaces of $\mathbb{R}^p$, rather than particular matrices spanning them. Accordingly, we are not so much concerned with R, and H themselves as with the following functions $\mathcal{R}_{(y,x)}$, and $\mathcal{H}$ of a general subspace $\mathcal{S}$ which they induce. By Proposition 1, we may define: $\mathcal{R}_{(y,x)}(\mathcal{S}) := R(y; u^T x)$, and $\mathcal{H}(\mathcal{S}) := H(u)$, where u is any matrix whose span is $\mathcal{S}$.

There are clear links with departures from independence for Hellinger integral of order two. Globally, $Y \perp\!\!\!\perp u^T X$ if and only if $R(y; u^T x) = 1$ for every supported $(y, u^T x)$, departures from unity at a particular $(y, u^T x)$ indicating local dependence between Y and $u^T X$. Moreover, noting that $\mathbb{E}\left[\left\{R(Y; u^T X)\right\}^{-1}\right] = 1$, we have $H(u) - 1 = \mathbb{E}\left[\frac{\left\{R(Y;u^T X)-1\right\}^2}{R(Y;u^T X)}\right]$. Thus, $H(u) - 1 \geq 0$, equality holding if and only if $Y \perp\!\!\!\perp u^T X$. Hence, we have the following result:

Proposition 2 *For any subspace $\mathcal{S}$ of $\mathbb{R}^p$,*

$$\mathcal{H}(\mathcal{S}) - \mathcal{H}(\{0_p\}) \geq 0,$$

where equality holds if and only if $Y \perp\!\!\!\perp X_{\mathcal{S}}$, and $\mathcal{H}(\{0_p\}) = 1$.

Since the rank of a matrix is the dimension of its span, there is no loss in requiring now that u is either 0_p or has full column rank d for some $1 \leq d \leq p$. Proposition 2 can be generalized from $(\{0_p\}, \mathcal{S})$ to any pair of nested subspaces $(\mathcal{S}_1, \mathcal{S}_1 \oplus \mathcal{S}_2)$, with $\mathcal{S}_1$ and $\mathcal{S}_2$ meeting only at the origin. We state the result below.

Proposition 3 *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces of $\mathbb{R}^p$ meeting only at the origin. Then,*

$$\mathcal{H}(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{H}(\mathcal{S}_1) \geq 0,$$

where equality holds if and only if $Y \perp\!\!\!\perp X_{\mathcal{S}_2} | X_{\mathcal{S}_1}$.

The above results establish $\mathcal{H}(\mathcal{S})$ as a natural measure of the amount of information on the regression of Y on X contained in a subspace $\mathcal{S}$, being strictly increasing with $\mathcal{S}$ except only when, conditionally on the dependence information already contained, additional dimensions carry no additional information. This property of Hellinger integral of order two can help establish the link with sufficient dimension reduction subspaces as we shall discuss in the next section.

2.3 Links with dimension reduction subspaces

The following result shows how we use Hellinger integral of order two to characterize dimension reduction subspaces and, thereby, the central subspace $\mathcal{S}_{Y|X} = \text{Span}(\eta)$ say, where η has full column rank $d_{Y|X}$.

Theorem 4 *We have:*

1. *$\mathcal{H}(\mathcal{S}) \leq \mathcal{H}(\mathbb{R}^p)$ for every subspace $\mathcal{S}$ of $\mathbb{R}^p$, equality holding if and only if $\mathcal{S}$ is a dimension reduction subspace (that is, if and only if $\mathcal{S} \supseteq \mathcal{S}_{Y|X}$).*

2. *All dimension reduction subspaces contain the same, full, regression information $H(I_p) = H(\eta)$, the central subspace being the smallest dimension subspace with this property.*

3. *$\mathcal{S}_{Y|X}$ uniquely maximizes $\mathcal{H}(\cdot)$ over all subspaces of dimension $d_{Y|X}$.*

The characterization of the central subspace given in the final part of Theorem 4 motivates consideration of the following set of maximization problems, indexed by the possible values d of $d_{Y|X}$. For each $d = 0, 1, ..., p$, we define a corresponding set of fixed matrices $\mathcal{U}_d$, whose members we call d-orthonormal, as follows:

$$\mathcal{U}_0 = \{0_p\} \text{ and, for } d > 0,\ \mathcal{U}_d = \left\{\text{all } p \times d \text{ matrices } u \text{ with } u^T u = I_d\right\}.$$

Noting that, for $d > 0$, u_1 and u_2 in $\mathcal{U}_d$ span the same d-dimensional subspace if and only if $u_2 = u_1 Q$ for some $d \times d$ orthogonal matrix Q. Since H is continuous and $\mathcal{U}_d$ is compact, there is an η_d maximizing $H(\cdot)$ over $\mathcal{U}_d$, so that $\text{Span}(\eta_d)$ maximizes $\mathcal{H}(\mathcal{S})$ over all subspaces of dimension d. And $\text{Span}(\eta_d)$ *is* unique when $d = d_{Y|X}$ (and, trivially, when $d = 0$). Putting

$$\overline{H}_d = \max\{H(u) : u \in \mathcal{U}_d\} = \max\{\mathcal{H}(\mathcal{S}) : \dim(\mathcal{S}) = d\}$$

and

$$\begin{aligned}\mathbb{S}_d &= \left\{\text{Span}(\eta_d) : \eta_d \in \mathcal{U}_d \text{ and } H(\eta_d) = \overline{H}_d\right\} \\ &= \left\{\mathcal{S} : \dim(\mathcal{S}) = d \text{ and } \mathcal{H}(\mathcal{S}) = \overline{H}_d\right\},\end{aligned}$$

Proposition 2 and Theorem 4 immediately give the following results.

Corollary 5 *In the above notation,*

1. $d > d_{Y|X} \Rightarrow \left[\overline{H}_d = H(I_p) \text{ and } \mathbb{S}_d = \left\{\mathcal{S} : \dim(\mathcal{S}) = d \text{ and } \mathcal{S} \supset \mathcal{S}_{Y|X}\right\}\right]$.

2. $d = d_{Y|X} \Rightarrow \left[\overline{H}_d = H(I_p) \text{ and } \mathbb{S}_d = \left\{\mathcal{S}_{Y|X}\right\}\right]$.

3. $d < d_{Y|X} \Rightarrow \overline{H}_d < H(I_p)$.

4. $d = 0 \Rightarrow \left[\overline{H}_d = 1 \text{ and } \mathbb{S}_d = \{0_p\}\right]$.

Furthermore, we have:

Proposition 6 $d_1 < d_2 \leq d_{Y|X} \Rightarrow 1 \leq \overline{H}_{d_1} < \overline{H}_{d_2}$.

The above results have useful implications for estimating the central subspace. In the usual case where $d_{Y|X}$ is unknown, they motivate seeking an H-optimal η_d for increasing dimensions d until $d = d_{Y|X}$ can be inferred. In Section 3, we will discuss such implications, and how they can help us to propose our method and an efficient computational algorithm.

2.4 From global to local

Having established the relation between H and the central subspace in previous section, naturally we need to propose an estimation method of H assuming known $d_{Y|X}$, and then an estimation procedure for $d_{Y|X}$. Directly estimating H involves density estimation, which can be overcome by kernel smoothing. We discover the link between H and three existing methods: (a) kernel discriminant analysis for categorical Y, as developed by Hernanedez and Velilla (2005), (b) sliced regression (Wang and Xia 2008), and (c) density minimum average variance estimation (Xia 2007). All three methods can be unified as adopting differently weighted H. More details are given in Section 6.2 of Appendix. The use of local kernel smoothing in the existing methods generally leads to accurate estimation, however, the computational burden increases very fast with the increase of sample size and predictor dimension. In this article, we propose a new

local approach with both estimation accuracy and computation efficiency in mind. We establish the link between local and global dependence on the central subspace, which guarantees that our local search can help find the (global) central subspace. The detailed discussion is provided in Section 6.3 of Appendix.

3 Estimation procedure

We directly approximate the Hellinger integral of order two via a local approach. Although similar in spirit to what developed by Xia (2007) and Wang and Xia (2008), rather than localize X, our approach taken here localizes (X, Y). This brings a number of benefits, including efficient computation, robustness and better handling of cases where Y takes only a few discrete values.

3.1 Weighted approximation

3.1.1 When response is continuous ...

To use the Hellinger integral of order two locally, we need to maximize $\frac{p(\eta^T x, y)}{p(\eta^T x)p(y)}$. Hence, the estimation of $p(\cdot)$ is critical. For convenience of derivation and without loss of generality, we assume $d_{Y|X} = 1$, suppress η, and consider a particular point (x_0, y_0): $\frac{p(x_0,y_0)}{p(x_0)p(y_0)}$. Note that centering $(x_0, y_0) = (0, 0)$ does not change the structure of the relations between x and y. That is, $\mathcal{S}_{Y-y_0|X-x_0} = \mathcal{S}_{Y|X}$. Hence, without loss of generality, we may further assume that $(x_0, y_0) = (0, 0)$. Let $w_0(x) = \frac{1}{h_1}K(\frac{x-x_0}{h_1})$, and $w_0(x, y) := \frac{1}{h_1}K(\frac{x-x_0}{h_1})\frac{1}{h_2}K(\frac{y-y_0}{h_2})$, where $K(\cdot)$ is a smooth kernel function, symmetric about 0, h_1 and h_2 being corresponding bandwidths. In all the following derivations, we assume all the density functions are differentiable up to the 4^{th} order and the smooth parameters follow the standard density estimation practice. More details can be found

in Jones (1996). Let $s_2 = \int u^2 K(u)du$ and

$$a_i = \int \frac{1}{h_1} K(\frac{x - x_0}{h_1}) x^i p(x)dx = \int K(u)(x_0 + h_1 u)^i p(x_0 + h_1 u)du,$$

we then have

$$\begin{aligned}
\mathbb{E}w_0(x) &= a_0 = p(x_0) + \frac{h_1^2}{2} s_2 p''(x_0) + O(h_1^4), \\
\mathbb{E}w_0(x)x &= a_1 = x_0 p(x_0) + h_1^2 s_2 p'(x_0) + \frac{h_1^2}{2} x_0 s_2 p''(x_0) + O(h_1^4), \quad \text{and} \\
\mathbb{E}w_0(x)x^2 &= a_2 = x_0^2 p(x_0) + 2h_1^2 s_2 x_0 p'(x_0) + h_1^2 s_2 p(x_0) + \frac{h_1^2}{2} x_0^2 s_2 p''(x_0) + O(h_1^4).
\end{aligned}$$

Hence,

$$\mathbb{E}w_0(x)x^2 - (\mathbb{E}w_0(x)x)^2 / \mathbb{E}w_0(x) = h_1^2 s_2 p(x_0) + O(h_1^4) \sim h_1^2 s_2 p(x_0). \tag{3.1}$$

Or, with centered $(x_0, y_0) = (0, 0)$, we simply have

$$\mathbb{E}w_0(x)x^2 = h_1^2 s_2 p(x_0) + O(h_1^4) \sim h_1^2 s_2 p(x_0). \tag{3.2}$$

Similarly,

$$\mathbb{E}w_0(y)y^2 - (\mathbb{E}w_0(y)y)^2 / \mathbb{E}w_0(y) = h_2^2 s_2 p(y_0) + O(h_1^4) \sim h_2^2 s_2 p(y_0), \tag{3.3}$$

or,

$$\mathbb{E}w_0(y)y^2 = h_2^2 s_2 p(y_0) + O(h_2^4) \sim h_2^2 s_2 p(y_0). \tag{3.4}$$

Secondly, let $p^{ij} = p^{ij}(x_0, y_0)$ be the corresponding partial derivatives of the density and $s_4 := \int u^4 K(u)du$. We have

$$\begin{aligned}
b_{ij} &= \int \frac{1}{h_1 h_2} K(\frac{x - x_0}{h_1}) K(\frac{y - y_0}{h_2}) x^i y^j p(x, y)dxdy \\
&= \int K(u)K(v)(x_0 + h_1 u)^i (y_0 + h_2 v)^j p(x_0 + h_1 u, y_0 + h_2 v)dudv
\end{aligned}$$

Hence,

$$\mathbb{E}w_0(x,y) = b_{00} = p(x_0,y_0) + \frac{h_1^2}{2}s_2p^{20} + \frac{h_2^2}{2}s_2p^{02} + \frac{h_1^2h_2^2}{4}s_2^2p^{22} + \frac{h_1^4}{4!}s_4p^{40} + \frac{h_2^4}{4!}s_4p^{04} + o(h_1^4+h_2^4), \tag{3.5}$$

$$\begin{aligned}\mathbb{E}w_0(x,y)xy = b_{11} &= x_0y_0p(x_0,y_0) + h_1^2s_2y_0p^{10} + h_2^2s_2x_0p^{01} + h_1^2h_2^2s_2^2p^{11} + \frac{h_1^2}{2}x_0y_0s_2p^{20} + \frac{h_2^2}{2}x_0y_0s_2p^{02}\\ &+ \frac{h_1^2h_2^2}{4}s_2^2x_0y_0p^{22} + \frac{h_1^2h_2^2}{2}s_2^2y_0p^{12} + \frac{h_1^2h_2^2}{2}s_2^2x_0p^{21} + \frac{h_2^4}{6}s_4x_0p^{03} + \frac{h_1^4}{6}s_4y_0p^{30}\\ &+ \frac{h_1^4}{4!}s_4x_0y_0p^{40} + \frac{h_2^4}{4!}s_4x_0y_0p^{04} + o(h_1^4+h_2^4),\end{aligned} \tag{3.6}$$

$$\begin{aligned}\mathbb{E}w_0(x,y)x^2y^2 = b_{22} &= x_0^2y_0^2[p(x_0,y_0) + \frac{1}{2}h_1^2s_2p^{20} + \frac{1}{2}h_2^2s_2p^{02} + \frac{1}{4}h_1^2h_2^2s_2^2p^{22} + \frac{1}{4!}h_1^4s_4p^{40} + \frac{1}{4!}h_2^4s_4p^{04}]\\ &+ x_0^2[h_2^2s_2p(x_0,y_0) + \frac{1}{2}h_1^2h_2^2s_2^2p^{20} + \frac{1}{2}h_2^4s_4p^{02}] + 2x_0h_1^2h_2^2s_2^2p^{10}\\ &+ y_0^2[h_1^2s_2p(x_0,y_0) + \frac{1}{2}h_1^2h_2^2s_2^2p^{02} + \frac{1}{2}h_1^4s_4p^{20}] + 2y_0h_1^2h_2^2s_2^2p^{01}\\ &+ 2x_0^2y_0[h_2^2s_2p^{01} + \frac{1}{3!}h_2^4s_4p^{03} + \frac{1}{2}h_1^2h_2^2s_2^2p^{21}] + h_1^2h_2^2s_2^2p(x_0,y_0)\\ &+ 2x_0y_0^2(h_1^2s_2p^{10} + \frac{1}{3!}h_1^4s_4p^{30} + \frac{1}{2}h_1^2h_2^2s_2^2p^{12}) + o(h_1^4+h_2^4).\end{aligned} \tag{3.7}$$

Together with centering of $(x_0,y_0)=(0,0)$, we have

$$\begin{aligned}&\mathbb{E}w_0(x,y)x^2y^2 - (\mathbb{E}w_0(x,y)xy)^2/\mathbb{E}w_0(x,y)\\ &= h_1^2h_2^2s_2^2p(x_0,y_0) + o(h_1^4+h_2^4) \sim h_1^2h_2^2s_2^2p(x_0,y_0),\end{aligned} \tag{3.8}$$

or

$$\mathbb{E}w_0(x,y)x^2y^2 = h_1^2h_2^2s_2^2p(x_0,y_0) + o(h_1^4+h_2^4) \sim h_1^2h_2^2s_2^2p(x_0,y_0), \tag{3.9}$$

Combining the above (3.1), (3.3) and (3.8), or, (3.2), (3.4) and (3.9), we then conclude that

$$\frac{p(x_0,y_0)}{p(x_0)p(y_0)} \sim H_0^*$$

where H_0^* is

$$\frac{\mathbb{E}w_0(x,y)x^2y^2 - (\mathbb{E}w_0(x,y)xy)^2/\mathbb{E}w_0(x,y)}{[\mathbb{E}w_0(x)x^2 - (\mathbb{E}w_0(x)x)^2/\mathbb{E}w_0(x)][\mathbb{E}w_0(y)y^2 - (\mathbb{E}w_0(y)y)^2/\mathbb{E}w_0(y)]},$$
$$\text{or, } \frac{\mathbb{E}w_0(x,y)x^2y^2}{\mathbb{E}w_0(x)x^2\mathbb{E}w_0(y)y^2}. \tag{3.10}$$

The above results have a very interesting interpretation, similar to

$$\frac{\text{'local variance ' of } x_0y_0}{(\text{'local variance ' of } x_0)(\text{'local variance ' of } y_0)} = \frac{V_0(xy)}{V_0(x)V_0(y)},$$

where V_0 is defined 'variance' at local of (x_0, y_0). That is, both have the 'correlation ' type formulation, with the latter one being non-centered. Thus finding η is basically, finding the best projection of X to maximize H_0^*. Either formula will suggest a SVD-type solution, avoiding the density estimation. The solution of η will be the principal eigenvector of

$$V_0(X)^{-1}V_0(XY)V_0(Y)^{-1}.$$

In particular, we will take KNN as the weight function for a fast computational algorithm.

3.1.2 When response is multivariate ...

The derivation in section 3.1.1 is based on a univariate response. One may have different ways to derive a formula for multivariate response, such as the projective re-sampling idea of Li, Wen and Zhu (2008). In this section, we simply note that the response appears in (3.10) is y^2, which can be considered as y^Ty when y is a vector. Thus a simple formula for multivariate response we shall use is

$$V_0(X)^{-1}V_0(XY^T)V_0(Y^T)^{-1},$$

where the principal eigenvector is again our solution.

3.1.3 When response is categorical ...

Similar argument in Section 3.1.1 leads to

$$\mathbb{E}(w_0(x)x^2|y=y_0) - (\mathbb{E}(w_0(x)x|y=y_0))^2/\mathbb{E}(w_0(x)|y=y_0) \sim h_2^2 s_2 p(x_0|y=y_0), \quad (3.11)$$

or,

$$\mathbb{E}(w_0(x)x^2|y=y_0) = h_2^2 s_2 p(x_0|y=y_0) + O(h_2^4) \sim h_2^2 s_2 p(x_0|y=y_0). \quad (3.12)$$

So, we have

$$\frac{p(x_0|y_0)}{p(x_0)} \sim \frac{V_0(x|y)h_1^2}{V_0(x)h_2^2}. \quad (3.13)$$

When KNN is used, based on the relationship between a KNN estimator and a kernel estimator with the same approximate bias and variance (Silverman 1986, page 99; Härdle et al 2004, page 101), we have $k_1 = 2nh_1 p(x_0)$, and $k_2 = 2nh_2 p(x_0|y=y_0)$. That is,

$$\frac{h_1^2}{h_2^2} = \left(\frac{p(x_0|y_0)/k_2}{p(x_0)/k_1}\right)^2, \quad (3.14)$$

where k_1 and k_2 are the sizes of the neighborhoods in estimating $p(x_0)$ and $p(x_0|y_0)$. Thus for fixed k_1 and k_2, putting (3.14) into (3.13), we have that $\frac{p(x_0|y_0)}{p(x_0)} \sim H_0^*$, where $H_0^* = \frac{V(x_0)}{V(x_0|y=y_0)}$. Hence, we want to find the principal eigenvector of

$$V_0(X|Y)^{-1}V_0(X).$$

3.2 Algorithm

For each observation (X_i, Y_i), we calculate the local dependence based on the development for the three cases in the previous sections, respectively.

(a) *Continuous univariate response Y:*

$$H_i^*(k) := V_{ki}(X)^{-1}V_{ki}(XY)V_{ki}(Y)^{-1}$$

where a subscript 'ki' denotes computation over the KNN of (X_i, Y_i).

(b) *Multivariate response Y:*

$$H_i^*(k) := V_{ki}(X)^{-1} V_{ki}(XY^T) V_{ki}(Y^T)^{-1}$$

The subscript 'ki' denote computation over the KNN of (X_i, Y_i).

(c) *Categorical univariate response* $Y \in \{1, ..., C\}$*:*

$$H_i^*(k) := V_{ki}(X|Y = j)^{-1} V_{ki}(X)$$

where, the subscript 'ki' here denotes computation over the KNN of X_i. Practically, a threshold on $p_{ki}(Y = j)$, the proportion of the observations from category j on the KNN of X_i, is used to guarantee the discriminatory power. An extreme case would be to discard $H_i^*(k)$ if none of the k Y values in the neighborhood differs from Y_i.

Assuming $d_{Y|X} = d$ is known and that $\{(X_i,\ Y_i),\ i = 1,\ 2,\ \cdots,\ n\}$ is an i.i.d. sample of (X, Y) values. Our estimation algorithm can be summarized as follows.

1. For each observation (X_i, Y_i), find its KNN in terms of the Euclidean distance $\|(X, Y) - (X_i, Y_i)\|$ ($\|X - X_i\|$, if Y is categorical) and η_i, the dominant eigenvector of $H_i^*(k)$.

2. Calculate the spectral decomposition of $\hat{M} := \frac{1}{n}\sum_{i=1}^n \eta_i \eta_i^T$, using its dominant d eigenvectors $\hat{u} := (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_d)$ to form an estimated basis of $\mathcal{S}_{Y|X}$.

The tuning parameter k plays a similar role to the bandwidth in nonparametric smoothing. Essentially, its choice involves a trade-off between estimation accuracy and exhaustiveness: for a large enough sample, a larger k can help improve the accuracy of estimated directions, while a smaller k increases the chance to estimate the central subspace exhaustively. In all numerical studies, a rough choice of k around $2p \sim 4p$

seemed working well. A larger k might be needed in the models with categorical response. More refined ways to choose k, such as cross-validation, could be used at greater computational expenses.

For the two versions of 'variance', in this paper we report the results of the non-centered version only. The two approaches give very similar results when the sample size is large. But the non-centered version has better performance for small and moderate sample size. This can be explained by less lower order terms in the non-centered version of approximation.

Section 2.2 shows that theoretically, the scale of predictor would not make any difference. However, practically when come to KNN, scale may make difference as neighborhood may be different under different scales. We find that using U-scale, $U = V^{1/2}Z$, in the intermediate steps seems the best and most consistent in our study, where $V = \text{diag}(\sigma_i)$ and σ_i is the variance of ith variable of X, $Z = \Sigma_X^{-1/2}[X - E(X)]$, and Σ_X is the covariance matrix of X.

Finally, in practice d is typically unknown, we shall propose an estimation method in the next section.

3.3 Determination of the structural dimension $d_{Y|X}$

Recall that $d_{Y|X} = 0$ is equivalent to $Y \perp\!\!\!\perp X$. At the population level, the eigenvectors of the kernel dimension reduction matrix, M say, represents a rotation of the canonical axes of $\mathbb{R}^p$ – one for each regressor – to new axes, with its eigenvalues reflecting the magnitude of dependence between Y and the corresponding regressors $\beta^T X$. At the sample level $\hat{M}$, holding the observed responses $\mathbf{y} := (y_1, ..., y_n)^T$ fixed while randomly permuting the rows of the $n \times p$ matrix $\mathbf{X} := (x_1, ..., x_n)^T$ will change $\hat{M}$ and tend to reduce the magnitude of the dependence – except when $d_{Y|X} = 0$.

Generally, consider testing H_0: $d_{Y|X} = m$ against H_a: $d_{Y|X} \geq (m + 1)$, for given

$m \in \{0, \cdots, p-1\}$. Let $B_m := (\hat{\beta}_1, \cdots, \hat{\beta}_m)$ and $A_m := (\hat{\beta}_{m+1}, \cdots, \hat{\beta}_p)$. The sampling variability in (B_m, A_m) apart, this is equivalent to testing $Y \perp\!\!\!\perp A_m^T X | B_m^T X$. Accordingly, the following procedure can be used to determine $d_{Y|X}$:

- Obtain $\hat{M}$ from the original $n \times (p+1)$ data matrix $(\mathbf{X}, \mathbf{y})$, compute its spectrum $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p$ and the test statistic

$$f_0 = \hat{\lambda}_{(m+1)} - \frac{1}{p-(m+1)} \sum_{i=m+2}^{p} \hat{\lambda}_i.$$

- Apply J independent random permutations to the rows of $\mathbf{X}A_m$ in the induced matrix $(\mathbf{X}B_m, \mathbf{X}A_m, \mathbf{y})$ to form J permuted data sets, obtain from each a new matrix $\hat{M}_j$ and a new test statistic f_j, $(j = 1, \cdots, J)$.

- Compute the permutation p-value:

$$p_{perm} := J^{-1} \sum_{j=1}^{J} I(f_j > f_0),$$

  and reject H_0 if $p_{perm} < \alpha$, where α is a pre-specified significance level.

- Repeat the previous three steps for $m = 0, 1, \ldots$ until H_0 cannot be rejected and take this m as the estimated $d_{Y|X}$.

3.4 Sparse version

In some applications, the regression model is held to have an intrinsic *sparse* structure. That is, only a few components of X affect the response. Then, effectively selecting informative predictors in the reduced directions can improve both estimation accuracy and interpretability. In this section, we incorporate a shrinkage estimation procedure proposed by Li and Yin (2008) into our method assuming d is known.

The central subspace $\mathcal{S}_{Y|X}$ is estimated by $\mathrm{Span}(\hat{u})$, where $\hat{u} := (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_d)$ are the d dominant eigenvectors of

$$\hat{M} := \frac{1}{n}\sum_{i=1}^{n} \eta_i \eta_i^T = \sum_{r=1}^{p} \hat{\lambda}_r \hat{\beta}_r \hat{\beta}_r^T \quad (\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p).$$

We begin by establishing that an alternative to arrive at this same estimate is to pool the $\{\eta_i\}$ by seeking $\tilde{u}$ with span as close as possible to $\{\mathrm{Span}(\eta_i)\}_{i=1}^n$ in the least-squares sense that

$$\tilde{u} := \underset{u \in \mathcal{U}_d}{\arg\min}\, g(u) \text{ where } g(u) := \sum\nolimits_{i=1}^{n} \left\| \eta_i - u\, u^T \eta_i \right\|^2, \tag{3.15}$$

with $u\, u^T$ being the orthogonal projector onto $\mathrm{Span}(u)$. The fact that $\mathrm{Span}(\tilde{u}) = \mathrm{Span}(\hat{u})$ now follows from observing that $g(u) = n - \sum_{r=1}^{p} \hat{\lambda}_r \hat{\beta}_r^T u\, u^T \hat{\beta}_r$ in which each $\hat{\beta}_r^T u\, u^T \hat{\beta}_r \leq 1$, with equality holding if and only if $\hat{\beta}_r \in \mathrm{Span}(u)$.

To select informative predictors, a shrinkage index vector α can be incorporated into this alternative formulation (3.15), as follows. With $\alpha \in R^p$ constrained by $\sum_{i=1}^{p} |\alpha_i| \leq \lambda$ for some $\lambda > 0$, let $\hat{\alpha}$ be the minimizer of

$$\sum_{i=1}^{n} ||\eta_i - \mathrm{diag}(\alpha)\tilde{u}\, \tilde{u}^T \eta_i||^2, \tag{3.16}$$

then, $\mathrm{diag}(\hat{\alpha})\tilde{u}$ forms a basis of the estimated sparse central space $\mathcal{S}_{Y|X}$. This constrained optimization (3.16) can be solved by a standard Lasso algorithm. Following Li and Yin (2008), we choose the tuning parameter λ by a modified Bayesian information criterion

$$BIC_\lambda = n \log\left(\frac{RSS_\lambda}{n}\right) + p_\lambda \log(nd),$$

where RSS_λ is the residual sum of squares from (3.16), and p_λ being the number of non-zero elements in $\hat{\alpha}$.

4 Evaluation

In this section, we evaluate the finite sample performance of the proposed method (H2 and sparseH2) through both simulation study and real data analysis. For comparison purposes, several existing methods (SIR, SAVE, PHD, MAVE and SR) were also evaluated in the simulation studies.

The matrix distance $\Delta(\hat{B}, B)$ was used to measure the estimation accuracy, where $\Delta(\hat{B}, B) = |\hat{B}(\hat{B}^T\hat{B})^{-1}\hat{B}^T - B(B^TB)^{-1}B^T|$ (Li, et al., 2005). Three sample sizes n=200, 400 and 600 were used in all numerical studies. The number of slices was fixed at either 5 (for n=200) or 10 (for n=400 and 600) for SIR, SAVE and SR when the response is continuous, otherwise the number of distinct Y values was used. Gaussian kernel and its corresponding optimal bandwidth were used for MAVE and SR. For each parameter setting, 200 data replicates were conducted.

4.1 Example 1: Estimation and Comparison.

In this example, we consider the following 4 models.

- Model I: $Y = (X^T\beta)^{-1} + 0.5\epsilon$,
- Model II: $Y = I[|X^T\beta_1 + 0.2\epsilon| < 1] + 2I[X^T\beta_2 + 0.2\epsilon > 0]$,
- Model III: $Y = 2(X^T\beta_1) + 2\exp(X^T\beta_2)\epsilon$,
- Model IV: $Y^* = 2(X^T\beta_1) + 2\exp(X^T\beta_2)\epsilon$, and $Y = 0,\ 1,\ 2$, for $Y^* \leq -2$, $-2 < Y^* < 2$ and $Y^* \geq 2$ respectively.

In all four models, $X \in \mathbb{R}^{10}$ is a 10-dimensional predictor, and ϵ is a standard normal noise which is independent of X. In model I and II, $X \sim N_{10}(0,\ \Sigma)$, with $\Sigma = \{\sigma_{ij}\} = \{0.5^{|i-j|}\}$. In model III and IV, $(x_1, \cdots, x_{10})$ are independently from a uniform distribution on $(-\sqrt{3}, \sqrt{3})$. In model I, $\beta = (1, 1, 1, 1, 0, \cdots, 0)^T$. In model II,

$\beta_1 = (1,1,1,1,0,\cdots,0)^T$ and $\beta_2 = (0,\cdots,0,1,1,1,1)^T$. While in models III and IV, $\beta_1 = (1,2,0,\cdots,0,2)^T/3$ and $\beta_2 = (0,0,3,4,0,\cdots,0)^T/5$.

Model I was studied by Wang and Xia (2008), where extreme values of Y occurs around the origin. Model II with discrete response {0, 1, 2, 3} was used by Zhu and Zeng (2006). Xia (2007) studied Model III, whose central subspace directions are in both regression mean and variance functions. Model IV is similar to Model III, except that the true response Y^* was not observable and only 3 class labels available. The results from 200 data replicates were reported in Table 1.

Table 1: Mean (standard deviation) of the estimation errors

	SIR	SAVE	PHD	MAVE	SR	H2
Model I						
$n = 200$	0.634(0.144)	0.734(0.169)	0.995(0.006)	0.986(0.049)	0.204(0.076)	0.400(0.122)
$n = 400$	0.493(0.112)	0.426(0.118)	0.996(0.005)	0.984(0.041)	0.114(0.037)	0.209(0.061)
$n = 600$	0.417(0.093)	0.331(0.093)	0.997(0.005)	0.984(0.043)	0.089(0.023)	0.173(0.049)
Model II						
$n = 200$	0.989(0.015)	0.580(0.194)	0.974(0.043)	0.318(0.158)	0.736(0.249)	0.392(0.097)
$n = 400$	0.985(0.022)	0.289(0.073)	0.971(0.044)	0.178(0.038)	0.365(0.185)	0.292(0.068)
$n = 600$	0.984(0.023)	0.205(0.043)	0.966(0.061)	0.142(0.028)	0.238(0.088)	0.234(0.056)
Model III						
$n = 200$	0.392(0.088)	0.805(0.171)	0.953(0.062)	0.747(0.164)	0.383(0.114)	0.394(0.112)
$n = 400$	0.266(0.056)	0.486(0.159)	0.954(0.058)	0.713(0.172)	0.231(0.054)	0.241(0.062)
$n = 600$	0.213(0.042)	0.429(0.140)	0.944(0.067)	0.664(0.169)	0.187(0.045)	0.201(0.049)
Model IV						
$n = 200$	0.464(0.104)	0.792(0.183)	0.946(0.069)	0.841(0.144)	0.628(0.166)	0.478(0.133)
$n = 400$	0.270(0.061)	0.631(0.196)	0.947(0.067)	0.742(0.171)	0.418(0.142)	0.327(0.078)
$n = 600$	0.216(0.044)	0.399(0.142)	0.949(0.071)	0.664(0.168)	0.329(0.079)	0.276(0.062)

The overall performance of the proposed H2 method is comparable to that of SR, with improvement in Model II and IV where the responses were discrete. SIR missed the symmetric pattern in model II, SAVE was sensitive to the number of slices and tended to miss the linear trend, MAVE focused on the regression mean function only and was not robust to the extreme values occurred in the response variable as in Model I. As claimed in the original paper (Wang and Xia 2008), SR performed well in all the models, especially with continuous responses. But our experience shows the estimation accuracy of SR can be affected by the number of distinct Y values in a categorical response model, especially when the number of categories is small. Furthermore, because of the use of local smoothing, the computation cost of SR increases exponentially with the increase of n. Table 2 gave a comparison of the computing time of SR and H2 methods for the above models. All the computation was done in Matlab version 7.12 on an office PC. Clearly we can see the advantage of proposed H2 method over SR, especially with the increase of sample size.

Table 2: Computation cost (CPU time in seconds) for 200 data replicates

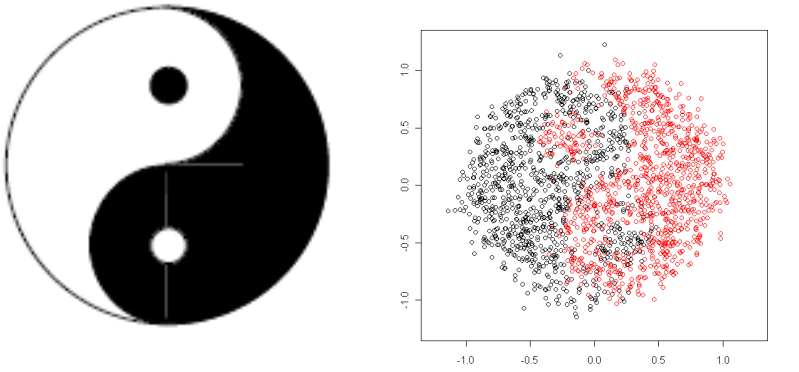
	Model I		Model II		Model III		Model IV	
	H2	SR	H2	SR	H2	SR	H2	SR
$n = 200$	11	457	10	416	10	397	10	445
$n = 400$	28	1224	27	1147	26	1185	31	1325
$n = 600$	44	1913	54	1940	43	2045	55	2078

4.2 Example 2: Tai Chi.

Consider the well-known *Tai Chi* figure in Asian culture shown in the left-hand panel of Figure 1. It is formed by one large circle, two medium half circles and two small circles. The regions with different colors are called *Ying* and *Yang*, respectively. They represent all kinds of opposite forces and creatures, yet work with each other with

harmony. Statistically, it is a difficult discrimination problem to separate them.

Figure 1: Tai Chi figure



(a) Tai Chi Figure

(b) Simulated Tai Chi model with 1000 observations and the projection onto first two H2 directions

Following Li (2000), we generate a binary regression data set with 10 covariates as follows: (1) x_1 and x_2 are the horizontal and vertical coordinates of points uniformly distributed within the large unit circle, the categorical response labels 1 and 2 being assigned to those located in the *Ying* and *Yang* regions respectively; (2) independently of this, $\{x_3, ..., x_{10}\}$ are i.i.d. standard normal random variables.

Li (2000) analyzed this example from the perspective of dimension reduction. Due to the binary response, SIR can only find 1 direction and so he proposed a double slicing scheme to identify the second direction. Here, we apply our method and SR to this model. Both our permutation test and cross-validation procedure in SR indicated a structural dimension of two. Table 3 shows that our H2 method outperforms SR, in which SR largely missed the second direction in the central subspace. The CPU time

(in seconds) again shows the efficiency of H2 approach.

Table 3: Tai Chi model with 200 data replicates

	$n = 200$		$n = 400$		$n = 600$	
	$\Delta(\hat{B}, B)$	CPU time	$\Delta(\hat{B}, B)$	CPU time	$\Delta(\hat{B}, B)$	CPU time
H2	0.443(0.163)	11	0.299(0.085)	29	0.237(0.066)	56
SR	0.836(0.171)	406	0.842(0.173)	1075	0.818(0.171)	2123

4.3 Example 3: Variable selection.

In this example, we evaluate the performance of proposed sparse version of H2 method (sparseH2) using model I and model III. To measure the effectiveness of variable selection, we use the true positive rate (TPR), defined as the ratio of the number of predictors correctly identified as active to the number of active predictors, and the false positive rate (FPR), defined as the ratio of the number of predictors falsely identified as active to the number of inactive predictors. Ideally, we wish to have TPR to be close to 1 and FPR to be close to 0 simultaneously. From Table 4, we can see the shrinkage procedure can effectively select informative covariates, and thus improve the estimation accuracy. With the increase of sample size, the estimation error and its standard deviation decrease.

Table 4: Effectiveness of variable selection

	Model I			Model III		
n	$\Delta(\hat{B}, B)$	TPR	FPR	$\Delta(\hat{B}, B)$	TPR	FPR
200	0.181(0.080)	1.000	0.083	0.384(0.150)	0.984	0.350
400	0.085(0.035)	1.000	0.043	0.213(0.065)	1.000	0.250
600	0.052(0.022)	1.000	0.031	0.173(0.056)	1.000	0.130

4.4 Example 4: Determining structural dimension

Here, we report the finite-sample performance of our permutation test in estimating $d_{Y|X}$. The results in Table 5 are based on 200 data replicates with sample size $n = 400$ for models used in the Example 1. The significance level was set at $\alpha = 0.05$, while $J = 1000$ permutations were used. The numbers in boldface are the percentages of correctly identified $d_{Y|X}$. Overall, the estimation of structural dimension is very reasonable.

Table 5: Permutation test for $d_{Y|X}$

	Percentage of estimated dimension				
Model	$d = 0$	$d = 1$	$d = 2$	$d = 3$	$d = 4$
I	0	**0.97**	0.03	0	0
II	0	0.12	**0.88**	0.00	0
III	0	0.21	**0.70**	0.09	0
IV	0	0.28	**0.67**	0.05	0

4.5 Example 5: A multivariate response model

With $X \sim N(\mathbf{0}, I_{10})$, the multivariate response model used here is:

$$Y_1 = 1/(X^T\beta_1) + 0.5\epsilon_1, \quad Y_2 = 2\exp(X^T\beta_2)\epsilon_2,$$

$$Y_3 = \epsilon_3 \quad \text{and} \quad Y_4 = \epsilon_4,$$

where $\beta_1 = (1,1,1,1,0,0,0,0,0,0)^T$, $\beta_2 = (0,0,0,0,0,0,1,1,1,1)^T$ and $\epsilon \sim N_4(\mathbf{0}, \Delta)$, with Δ=diag(Δ_1, I_2), in which

$$\Delta_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},$$

so that $S_{Y|X} = \text{span}(\beta_1, \beta_2)$.

Table 6: The estimation errors

n	H2	PR-H2	PR-SIR
200	0.325(0.154)	0.306(0.138)	0.605(0.187)
400	0.144(0.040)	0.133(0.039)	0.418(0.127)
600	0.116(0.031)	0.101(0.029)	0.253(0.068)

SR method is not directly applicable here. We compare our H2 method for multivariate responses, our univariate H2 method together with the projective re-sampling (PR-H2) approach of Li, Wen and Zhu (2008), and projective re-sampling with sliced inverse regression (PR-SIR). Following Li, Wen and Zhu (2008), the numbers of slices used are 5, 10 and 10 in SIR corresponding to sample sizes of 200, 400 and 600, respectively. The Monte Carlo sample size is m_n=2000 for the PR type approaches. Numerical results are given in Table 6. Overall, H2 approach performed pretty well, giving better results than PR-SIR. As expected, the estimation accuracy of all methods improves with the increase of sample size.

4.6 Example 6: Communities and crime

There have been extensive studies on the relationship between violent crimes and the socio-economic environment. This data set contains information from three sources: the social-economic data from the 1990 US census, the law enforcement data from the 1990 US LEMAS survey and the crime data from the 1995 FBI UCR. Further details on the data and on the attributes used can be found at the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets.html`). There are $n = 1994$ observations from different communities across the US. The response variable is the per capita number of violent crimes. The predictors included in our analysis are shown in the second column of Table 7.

Table 7: Community and Crime

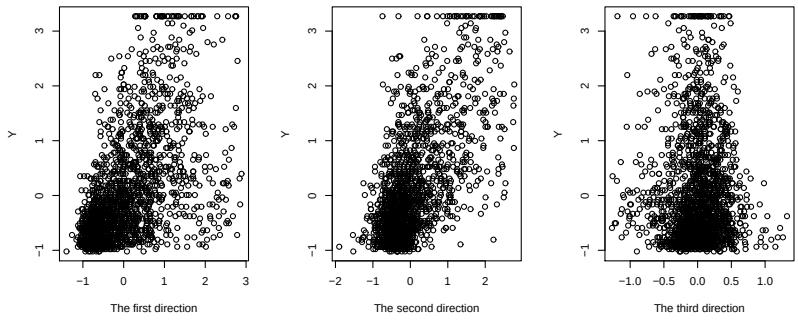
	Predictor	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
x_1	percentage of population that is 65 and over in age	-0.05	0.04	0.11
x_2	median family income	0.08	0.18	0.69
x_3	percentage of people under the poverty level	0.91	-0.20	0.15
x_4	unemployment rate	0.02	-0.01	0.19
x_5	percentage of population who are divorced	0.04	-0.01	-0.04
x_6	percentage of kids born to never married	0.17	0.93	-0.17
x_7	percentage of people who speak only English	-0.04	0.03	0.07
x_8	mean persons per household	-0.05	-0.02	0.08
x_9	percentage of people in owner occupied households	0.22	0.09	-0.34
x_{10}	percentage of housing occupied	0.05	-0.07	-0.05
x_{11}	median value of owner occupied house	0.16	-0.17	-0.53
x_{12}	population density in persons per square mile	0.15	0.02	0.02
x_{13}	percent of people using public transit for commuting	-0.13	-0.11	-0.05

All the variables were normalized into the range 0.00-1.00 using an unsupervised, equal-interval binning method. The distributions of most predictors are very skew, which precludes the use of inverse regression methods. The large sample size also prevents the use of SR.

In practice, real data sets such as this can have a low signal-to-noise ratio. In such cases, we have found it very helpful to filter out 'noisy neighbors' in both the estimation of directions and permutation test parts of our method. This is achieved straightforwardly by retaining only those cases for which, as a proportion of the total, the largest eigenvalue in each local exceeds a specified threshold. Here, we use 0.50 as the threshold value, the permutation test giving p-values of 0, 0, 0.014 and 0.328 for d=0, 1, 2, and 3.

With d=3, we find the direction estimates reported in Table 7. The first direction is dominated by x_3, the percentage of people under the poverty level, and the second by

Figure 2: Community and crime



x_6, the percentage of kids born to never married parents, while the third direction can be seen as a combination of variables related to family structure. The scatter plots of response against each of these three directions (Figure 2) confirm their importance. Both the poverty level and the percentage of kids with unmarried parents have a significant positive effect on the crime rate. The contrast between family income and house value is another important factor.

5 Discussion

In this paper, we propose a new and general approach to dimension reduction based on the Hellinger integral of order two, the underlying global-through-local theory endorsing its naturalness. Rather than localizing X as in Xia (2007) and Wang and Xia (2008), its implementation is on joint local in (X, Y). This brings a number of benefits, including robustness, computation efficiency and better handling of cases where Y takes only a few discrete values.

In summary, our approach has several advantages. It combines speed with minimal (essentially, just existence) assumptions, while performing well in terms of estimation accuracy, robustness and exhaustiveness, this last due to its local weighted approximation.

Relative to existing methods, examples show that our approach: (a) is computationally efficient without sacrificing estimation accuracy, allowing larger problems to be tackled, (b) is more general, multidimensional (discrete or continuous) Y being allowed, and (c) benefits from having a sparse version, this enabling variable selection while making overall performance broadly comparable. Finally, incorporating appropriate weights, it unifies three existing methods, including sliced regression, kernel discrimination and density MAVE.

Among other further work, a global search of the Hellinger integral of order two with or without slicing, similar to that of Hernánedez and Velilla (2005), merits investigation. That said, being based on nonparametric density estimation, this might improve estimation accuracy especially when $d_{Y|X} > 1$. However, this brings some other issues, such as the starting value and computation complexity.

6 Appendix: Additional materials

In this section, we provide additional technical details. First part provides proofs of some theoretical results. Second part establishes the connections between Hellinger integral of order two and some existing methods. The last part connects between central subspace and local central subspace.

6.1 Additional justifications

Proposition 1

$\text{Span}(u_1) = \text{Span}(u_2) \Rightarrow \text{rank}(u_1) = \text{rank}(u_2) = r$, say. Suppose first that $r = 0$. Then, for $i = 1, 2$, u_i vanishes, so that $Y \perp\!\!\!\perp u_i^T X$ implying $R(y; u_i^T x) \overset{(y,x)}{\equiv} 1$. Otherwise, let u be any matrix whose $1 \leq r \leq p$ columns form a basis for $\text{Span}(u_1) = \text{Span}(u_2)$. Then, for $i = 1, 2$, $u_i = uA_i^T$ for some A_i of full column rank, so that

$$u_1^T X = u_1^T x \Leftrightarrow u^T X = u^T x \Leftrightarrow u_2^T X = u_2^T x.$$

Thus, $p(y|u_1^T x) \overset{(y,x)}{\equiv} p(y|u_2^T x)$ implying $R(y; u_1^T x) \overset{(y,x)}{\equiv} R(y; u_2^T x)$, and $H(u_1) = H(u_2)$.

Proposition 3

Let $\mathcal{S}_1 = \text{Span}(u_1)$ and $\mathcal{S}_2 = \text{Span}(u_2)$ be nontrivial subspaces of $\mathbb{R}^p$ meeting only at the origin, so that (u_1, u_2) has full column rank and spans their direct sum $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{x_1 + x_2 : x_1 \in \mathcal{S}_1, x_2 \in \mathcal{S}_2\}$. Then, $\mathcal{H}(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{H}(\mathcal{S}_1)$ can be evaluated using conditional versions of $\mathcal{R}_{(y,x)}$ and $\mathcal{H}$, defined as follows.

We use $R(y; u_2^T x | u_1^T x)$ to denote the conditional dependence ratio:

$$\frac{p(y, u_2^T x | u_1^T x)}{p(y|u_1^T x) p(u_2^T x | u_1^T x)} = \frac{p(y|u_1^T x, u_2^T x)}{p(y|u_1^T x)} = \frac{p(u_2^T x | y, u_1^T x)}{p(u_2^T x | u_1^T x)},$$

so that $Y \perp\!\!\!\perp u_2^T X | u_1^T X$ if and only if $R(y; u_2^T x | u_1^T x) \overset{(y,x)}{\equiv} 1$, while

$$R(y; u_1^T x, u_2^T x) = R(y; u_1^T x) R(y; u_2^T x | u_1^T x). \tag{6.1}$$

Then, defining the conditional Hellinger integral of order two $H(u_2|u_1)$ by

$$H(u_2|u_1) := \mathbb{E}_{u_2^T X | (Y, u_1^T X)} \left\{ R(Y; u_2^T X | u_1^T X) \right\},$$

(6.1) gives:

$$H(u_1, u_2) = \mathbb{E}_{(Y,X)} \{ R(Y; u_1^T X) H(u_2|u_1) \}. \tag{6.2}$$

Noting that $\mathbb{E}_{u_2^T X | (Y, u_1^T X)} \left[\left\{ R(Y; u_2^T X | u_1^T X) \right\}^{-1} \right] = 1$, we have

$$\mathbb{E}_{(Y,X)} R(Y; u_1^T X) \left[\frac{\left\{ R(Y; u_2^T X | u_1^T X) - 1 \right\}^2}{R(Y; u_2^T X | u_1^T X)} \right] = \mathbb{E}_{(Y,X)} R(Y; u_1^T X) \{ H(u_2|u_1) - 1 \}$$

$$= H(u_1, u_2) - H(u_1) = \mathcal{H}(\mathcal{S}_2 \oplus \mathcal{S}_1) - \mathcal{H}(\mathcal{S}_1).$$

The last equality due to that $p(y|u_1^T x)$ and $p(y|u_1^T x, u_2^T x)$ do not depend on the choice of u_1 and u_2. We complete the proof.

Theorem 4

Since the central subspace is the intersection of all dimension reduction subspaces, it suffices to prove the first assertion. If $\mathcal{S} = \mathbb{R}^p$, the result is trivial. Again, if $\mathcal{S} = \{0_p\}$, it follows at once from Proposition 2. Otherwise, it follows from Proposition 3, taking $\mathcal{S}_2$ as the orthogonal complement in $\mathbb{R}^p$ of $\mathcal{S}_1 = \mathcal{S}$.

Proposition 6

The inequality $1 \leq \overline{H}_{d_1}$ is immediately from Proposition 2. The proof that $\overline{H}_{d_1} < \overline{H}_{d_2}$ is by contradiction. Consider $d_1 > 0$ and, for a given η_{d_1} such that $H(\eta_{d_1}) = \overline{H}_{d_1}$, let u be any matrix such that $(\eta_{d_1}, u) \in \mathcal{U}_{d_2}$. Then, based on Proposition 3

$$\overline{H}_{d_2} - \overline{H}_{d_1} \geq H(\eta_{d_1}, u) - H(\eta_{d_1}) \geq 0.$$

If $\overline{H}_{d_1} = \overline{H}_{d_2}$, then, $H(\eta_{d_1}, u) = H(\eta_{d_1})$ for any u. Again by Proposition 3, $Y \perp\!\!\!\perp u^T X | \eta_{d_1}^T X$. It follows that $\text{Span}(\eta_{d_1})$ is a dimension reduction subspace, contrary to $d_1 < d_{Y|X}$. The proof for $d_1 = 0$ follows from the same argument.

6.2 Unification of three existing methods

Kernel Discriminant Analysis (Hernánedez and Velilla 2005)

Suppose that Y is a discrete response where, for some countable index set $\mathcal{Y} \subset \mathbb{R}$, $Y = y$ with probability $p(y) > 0$ $\left(\sum_{y \in \mathcal{Y}} p(y) = 1\right)$ and, we assume, for each $y \in \mathcal{Y}$, X admits a conditional density $p(x|y)$ so that

$$p(x) = \sum_{y \in \mathcal{Y}} p(y, x) \text{ where } p(y, x) = p(y)p(x|y) = p(x)p(y|x),$$

whence

$$p(u^T x) = \sum_{y \in \mathcal{Y}} p(y, u^T x) \text{ where } p(y, u^T x) = p(y)p(u^T x|y) = p(u^T x)p(y|u^T x). \quad (6.3)$$

In the discrete case, we have

$$\begin{aligned} H(u) &= \mathbb{E}\left(\frac{p(u^T X|Y)}{p(u^T X)}\right) \\ &= \mathbb{E}_Y \mathbb{E}_{u^T X|Y}\left(\frac{p(u^T X|Y)}{p(u^T X)}\right) \\ &= \sum_{y\in\mathcal{Y}} p(y) \int \left(\frac{p^2(u^T x|y)}{p(u^T x)}\right). \end{aligned}$$

Hernánedez and Velilla (2005) proposed a method which maximises the following index:

$$I_{HV}(u) := \sum_{y\in\mathcal{Y}} var_{u^T X}\left(p(y)\frac{p(u^T x|y)}{p(u^T x)}\right).$$

Since

$$\begin{aligned} I_{HV}(u) &= \sum_{y\in\mathcal{Y}} p^2(y) var_{u^T X}\left(\frac{p(u^T x|y)}{p(u^T x)}\right) \\ &= \sum_{y\in\mathcal{Y}} p^2(y) \int \left(\frac{p^2(u^T x|y)}{p(u^T x)}\right) - a \end{aligned}$$

where $a := \sum_{y\in\mathcal{Y}} p^2(y)$ is constant, their index is equivalent to ours, except that the weight function $p(y)$ is squared.

Sliced Regression (Wang and Xia 2008).

Let Y be sliced into k slices, with C_i denoting the set of y values in the i^{th} slice. Then,

$$\mathbb{E}_{(X,Y)}\left(\frac{p(Y|X)}{p(Y)}\right) = \mathbb{E}_X \sum_{i=1}^{k}\left(\frac{(p(C_i|X))^2}{p(C_i)}\right) = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}_X\left\{\mathbb{E}^2_{Y|X}(I_{C_i}(Y)|X)\right\} \quad (6.4)$$

while, using $\mathbb{E}(I^2_{C_i}(Y)) = \mathbb{E}(I_{C_i}(Y)) = p(C_i)$, we have

$$k = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}(I^2_{C_i}(Y)) = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}_X\mathbb{E}_{Y|X}(I^2_{C_i}(Y)|X). \quad (6.5)$$

Denoting the sliced form of Y by $\tilde{Y}$, (6.4) and (6.5) together give

$$\begin{aligned} k - \mathbb{E}\left(\frac{p(Y|X)}{p(Y)}\right) &= \sum_{i=1}^{k} \frac{1}{p(C_i)} \mathbb{E}_X \left\{ \mathbb{E}_{Y|X}(I^2_{C_i}(Y)|X) - \mathbb{E}^2_{Y|X}(I_{C_i}(Y)|X) \right\} \\ &= \sum_{i=1}^{k} \frac{1}{p(C_i)} \mathbb{E}_X \left[\mathbb{E}_{Y|X}\{I_{C_i}(Y) - \mathbb{E}_{Y|X}(I_{C_i}(Y)|X)\}^2 |X) \right] \\ &= \mathbb{E}_X \mathbb{E}_{\tilde{Y}} \mathbb{E}_{Y|X} \left[\left\{ \left(\frac{I_{\tilde{Y}}(Y)}{p(\tilde{Y})} \right) - \mathbb{E}\left(\frac{I_{\tilde{Y}}(Y)}{p(\tilde{Y})} \right) |X \right\}^2 \right] \end{aligned}$$

so that, through slicing, optimizing the Hellinger integral of order two can be reformulated as weighted least squares estimation. Thus, any method for finding the dimensions in the mean function can be used. In particular, if the procedure of minimum average variance estimation (Xia, Tong, Li and Zhu, 2002) is used, we recover the sliced regression method of Wang and Xia (2008), apart from the weights $p(\tilde{Y})^{-2}$.

Density minimum average variance estimation (Xia 2007)

As in Fan, Yao and Tong (1996), the conditional density can be written as $p(y|x) = \mathbb{E}_{Y|x}(G_h(Y-y)|x)$, where G is a kernel and h is the bandwidth, so that

$$\mathbb{E}\left(\frac{p(Y|X)}{p(Y)}\right) = \int \frac{p(x)}{p(y)} \mathbb{E}^2_{Y|x}(G_h(Y-y)|x) dxdy.$$

Thus, defining the constant $a_0 := \int \frac{p(x)}{p(y)} \mathbb{E}_{Y|x} G^2_h(Y-y) dxdy$, we have

$$\begin{aligned} a_0 - \mathbb{E}\left(\frac{p(Y|X)}{p(Y)}\right) &= \int \frac{p(x)}{p(y)} \mathbb{E}\left\{G_h(Y-y) - \mathbb{E}(G_h(Y-y)|x)\right\}^2 dxdy \\ &= \int p(x)p(y) \mathbb{E}\left\{ \frac{G_h(Y-y)}{p(y)} - \mathbb{E}\left(\frac{G_h(Y-y)}{p(y)} |x \right) \right\}^2 dxdy \\ &= \mathbb{E}_x \mathbb{E}_y \mathbb{E}_{Y|x} \left\{ \frac{G_h(Y-y)}{p(y)} - \mathbb{E}\left(\frac{G_h(Y-y)}{p(y)} |x \right) \right\}^2 \end{aligned}$$

Therefore, dMAVE and dOPG developed by Xia (2007) are methods to estimate the last term, apart from the weight $p(y)^{-2}$.

6.3 Local central subspaces

In this section, we define a local central subspace. Let $\Omega := \{(x, y) : f(x, y) > 0\}$ denote the support of (X, Y), inducing the marginal supports:

$$\Omega_X := \{x : \exists y \text{ with } f(x, y) > 0\} \text{ and } \Omega_Y := \{y : \exists x \text{ with } f(x, y) > 0\}.$$

Let $(x, y) \in \Omega$, $L_x \subseteq \Omega_X$ and $L_y \subseteq \Omega_Y$ be neighborhoods of x and y, respectively. And let $L := L_X \times L_Y$. One can define a local dimension reduction subspace as a subspace spanned by B such that $Y \perp\!\!\!\perp X|B^T X$ for $(X, Y) \in L$. Hence, a local central subspace (LCS) is the intersection of all local dimension reduction subspaces and if the intersection itself, say, B_l, also satisfies $Y \perp\!\!\!\perp X|B_l^T X$ for $(X, Y) \in L$. For simplicity we denote LCS by $\mathcal{S}_{L(Y|X)}$. Note that if $L = \Omega$, then $\mathcal{S}_{L(Y|X)} = \mathcal{S}_{Y|X}$. Moreover, define $W = W(x, y) = 1$ if $(x, y) \in L$, and 0 otherwise, then $\mathcal{S}_{L(Y|X)} = \mathcal{S}_{Y|(X,W=1)}$, the CS conditionally on $W = 1$ (that is, within the subpopulation identified by $W = 1$; Chiaromonte, Cook and Li, 2002).

Immediately from the result in Section 2 that maximize Hellinger integral of order two will give us a basis of the CS, maximizing Hellinger integral of order two over L will give us a basis of the LCS. However, our main goal here is to establish the relations between CS and LCS.

Suppose that $f_l(y|x)$ is the local density for $(x, y) \in L$, and $f_l(y|x) = f_l(y|B_l^T x)$, where $B_l = (\beta_1, \beta_2, ..., \beta_q)$ whose columns form a basis of $\mathcal{S}_{L(Y|X)}$. Let $\frac{\partial}{\partial x} = (\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_p})^T$ denote the gradient operator, and $u = B_l^T x = (u_1, ..., u_q)^T$. Then, our first result is a characterization of the LCS.

Proposition 7 $\mathcal{S}_{L(Y|X)} = Span\{\frac{\partial}{\partial x} f_l(y|x), (x, y) \in L\}$.

An immediate application is to $L = \Omega$. Hence,

Corollary 8 $\mathcal{S}_{Y|X} = Span\{\frac{\partial}{\partial x} f(y|x), (x, y) \in \Omega\}$.

The above results extend the results in Zhu and Zeng (2006). The following proof is very similar to that of Zhu and Zeng (2006).

Proof of Proposition 7. It is sufficient to show that for any $\alpha \in \mathbb{R}^p$, $\alpha^T B_l = 0$ iff $\alpha^T \frac{\partial}{\partial x} f_l(y|x) = 0$ for all $(x, y) \in L$.

By the chain rule of differentiation, $\frac{\partial}{\partial x} f_l(y|x) = B_l \frac{\partial}{\partial u} f_l(y|u)$. Thus $\alpha^T B_l = 0$ implies that $\alpha^T \frac{\partial}{\partial x} f_l(y|x) = 0$ for all $(x, y) \in L$.

We show the other way around by contradiction. Assume that these exists $\alpha_0 \in \mathbb{R}^p$ such that $\alpha_0^T \frac{\partial}{\partial x} f_l(y|x) = 0$ for all $(x, y) \in L$, but $\alpha_0^T B_l = 0$. Then $\xi_1 = B_l^T \alpha_0 / ||B_l^T \alpha_0||$ is a nonzero $q \times 1$ vector. Hence, $\alpha_0^T \frac{\partial}{\partial x} f_l(y|x) = \alpha_0^T B_l \frac{\partial}{\partial u} f_l(y|u) = 0$ implies that $\xi_1^T \frac{\partial}{\partial u} f_l(y|u) = 0$, which means that the directional derivative of f_l as a function of u along ξ_1 is always 0. Hence, $f_l(y|u) = f_l(y; u)$ is a constant along ξ_1. That is, $f_l(y; u + t\xi_1) = f_l(y, u)$ for all $t \in \mathbb{R}$. We then expand ξ_1 to form an orthonormal basis for $\mathbb{R}^q$, say, $A = (\xi_1, ..., \xi_q)$. And define $v = A^T u = (v_1, ..., v_q)^T$, then $f_l(y; u) = f_l(y; Av)$, and $\frac{\partial}{\partial v_1} f_l(y; Av) = \xi_1^T \frac{\partial}{\partial u} f_l(y; u) = 0$. Thus $f_l(y; Av)$ doesn't depend on v_1, so we can write $f_l(y|u) = f_l(y; u) = f_l(y; Av) = \tilde{f}_l(y; v_2, ..., v_q) = \tilde{f}_l(y; \xi_2^T B_l^T x, ..., \xi_q^T B_l^T x)$. Therefore, $(B_l \xi_2, ..., B_l \xi_q)$ is a local dimension reduction subspace, which has structural dimension $q - 1$. This contradicts to that the LCS has dimension q. We complete the proof.

Next we establish relations between LCS and CS. For $(x, y) \in L$, we have

$$f_l(x, y) := f((x, y)|(X, Y) \in L) = \frac{f(x, y)}{P_l}, \tag{6.6}$$

where $P_l := \mathbb{P}((X, Y) \in L) = \int \int_L f(x, y) dx dy$. Let $f_{l_y}(x) = \int_{L_y} f(x, y) dy$, then

$$f_l(x) := f(x|(X, Y) \in L) = \frac{f_{l_y}(x, y)}{P_l}. \tag{6.7}$$

Define $w = w(y) = 1$ if $y \in L_y$, and 0, otherwise. We then have

$$f(w = 1|x) := \mathbb{E}_{(Y|X=x)} w(Y) = \mathbb{P}(w(y) = 1|x),$$

which means that

$$f(w = 1|x) = \int_{\Omega_Y} w(y) f(y|x) dy = \int_{L_y} f(y|x) dy = f_{l_y}(x)/f(x). \tag{6.8}$$

Combining (6.6), (6.7) and (6.8), we have

$$f_l(y|x) = \frac{f_l(x, y)}{f_l(x)} = \frac{f(x, y)}{f_{l_y}(x)} = \frac{f(y|x)}{f(w = 1|x)},$$

or equivalently,

$$f(y|x) = f_l(y|x) f(w = 1|x). \tag{6.9}$$

By taking derivatives, we have

$$\frac{\partial}{\partial x} f(y|x) = f(w = 1|x) \frac{\partial}{\partial x} f_l(y|x) + f_l(y|x) \frac{\partial}{\partial x} f(w = 1|x). \tag{6.10}$$

If $L_y = \Omega_Y$, that is, locality is only on x, then $f(w = 1|x) = 1$ a constant. Hence, (6.10) becomes $\frac{\partial}{\partial x} f(y|x) = \frac{\partial}{\partial x} f_l(y|x)$. Based on Proposition 7, we have that

Corollary 9 *$\mathcal{S}_{Y|X} = Span\{\mathcal{S}_{L(Y|X)},$ for all $L = L_x \times L_y,$ and $L_y = \Omega_Y\}$.*

If further assume that Ω_X is an open set, then Corollary 9 reduces to the result of Yin, Wang, Li and Tang (2010). Another direct application is when Y is categorical or discrete case, for which locality may be only meaningful on X, hence, $L_y = \Omega_Y$.

In general, however, note that $w(y)$ is a function of y, Hence, $f(w = 1|x) = f(w = 1|\eta^T x)$, where η is a basis matrix of CS. Then $\frac{\partial}{\partial x} f(w = 1|x) = \eta \frac{\partial}{\partial(\eta^T x)} f(w = 1|\eta^T x)$, together with Proposition 7 and (6.10) we have

$$\frac{\partial}{\partial x} f_l(y|x) = \frac{1}{f(w = 1|x)} [\frac{\partial}{\partial x} f(y|x) - f_l(y|x) \frac{\partial}{\partial x} f(w = 1|x)] \subseteq \mathcal{S}_{Y|X}. \tag{6.11}$$

Therefore,

Corollary 10 *Span*$\{\mathcal{S}_{L(Y|X)},$ *for all* $L = L_x \times L_y\} \subseteq \mathcal{S}_{Y|X}$.

However, practically we often expect the equality holds. In fact, we don't tempt or advise to estimate the dimensions in $f(w = 1|x)$. If a model is nice in a way that there is no 'extreme' values of Y, then for locals we have $L_y = \Omega_Y$ and the two are equal. If a model is not nice in a way to produce 'extreme' values of Y, then w is not a constant, however, the dimensions in $f(w = 1|x)$ are due to 'extreme' values for which they should not belong to CS. Our locality using $f_l(y|x)$ in fact is filtering these dimensions. Furthermore, if L_y doesn't depend on x (often a situation), then based on (6.8), we have that

$$\frac{\partial}{\partial x} f(w = 1|x) = \int_{L_y} \frac{\partial}{\partial x} f(y|x) dy = a_1 \frac{\partial}{\partial x} f(y|x) + a_2 \epsilon,$$

where ϵ is a unit length vector in $\mathcal{S}_{Y|X}$ but orthogonal to $\frac{\partial}{\partial x} f(y|x)$, while a_1 and a_2 are scalars with $a_1 = O(\delta_y)$ and $a_2 = O(\delta_y^2)$, and δ_y is the size of L_y. Based on (6.10),

$$[1 - a_1 f_l(y|x)] \frac{\partial}{\partial x} f(y|x) - a_2 f_l(y|x)\epsilon = f(w = 1|x) \frac{\partial}{\partial x} f_l(y|x). \tag{6.12}$$

Considering the magnitudes of a_1, a_2 in (6.12), approximately we often expect that in practice $\frac{\partial}{\partial x} f(y|x) \propto \frac{\partial}{\partial x} f_l(y|x)$. Hence, the two sides in Corollary 10 are expected to be the same.

References

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society* B, 63, 393–410.

Chiaromonte, F., Li, B. and Cook, R. D. (2002). Sufficient 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. (1998a). *Regression Graphics: Ideas for studying regressions through graphics*. Wiley: New York.

Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93, 84–100.

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.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83, 189-196.

Härdle, W. and Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and semiparametric models*. Springer: New York.

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

Hernández, A. and Velilla, S. (2005). Dimension reduction in nonparametric kernel discriminant analysis, *Journal of Computational and Graphical Statistics*, 14, 847–866.

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

Jones, M. C. (1996). The local dependence function. *Biometrika*, 83, 899-904.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, 33, 1580–1616.

Li, B., Wen, S. and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, 103, 1177–1186.

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.

Li, K. C. (2000). High dimensional data analysis via the SIR/PHD approach. Lecture Notes obtained from `http://www.stat.ucla.edu/~kcli/sir-PHD.pdf`

Li, L. and Yin, X. (2008). Sliced inverse regression with regularizations. *Biometrics*, 64, 124-131.

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

Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis.* Chapman & Hall/CRC: London.

Wang, H. and Xia, Y. (2008). Sliced Regression for Dimension Reduction. *Journal of the American Statistical Association*, 103, 811–821.

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

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

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. *Journal of the Royal Statistical Society* B, 64, 159–175.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika*, 90, 113-125.

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

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Annals of Statistics*, 39, 3392-3416.

Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a Multiple-index regression. *Journal of Multivariate Analysis*, 99, 1733–1757.

Yin, X., Wang, Q., Li, B. and Tang, Z. (2010). On aggregate dimension reduction and variable selection. *manuscript*.

Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101, 271-290.

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