Sufficient dimension reduction based on the Hellinger integral: a general, unifying approach

Xiangrong Yin* Frank Critchley† Qin Wang‡

June 7, 2010

Abstract

Sufficient dimension reduction provides a useful tool to study the dependence between a response Y and a multidimensional regressor X, sliced regression (Wang and Xia, 2008) being reported to have a range of advantages – estimation accuracy, exhaustiveness and robustness – over many other methods. A new formulation is proposed here based on the Hellinger integral of order two – and so jointly local in (X, Y) – together with an efficient estimation algorithm. The link between χ^2-divergence and dimension reduction subspaces is the key to our approach, which has a number of strengths. It requires minimal (essentially, just existence) assumptions. Relative to sliced regression, it is faster, allowing larger problems to be tackled, more general, multidimensional (discrete, continuous or mixed) Y as well as X being allowed, and includes a sparse version enabling variable selection, while overall performance is broadly comparable, sometimes better (especially, when Y takes only a few discrete values). Finally, it unifies three existing methods, each being shown to be equivalent to adopting suitably weighted forms of the Hellinger integral.

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

*Department of Statistics, The University of Georgia

†Department of Mathematics and Statistics, The Open University, UK

‡Department of Statistical Sciences and Operations Research, Virginia Commonwealth University

1 Introduction

In simple regression a 2D 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. Regression visualization, as conceived here, 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 $Y|X$, X, $X|Y$ or Y. Borrowing a phrase from classical statistics, the key idea is 'sufficient dimension reduction'. That is, reduction of the dimension of the predictors without loss of information on the conditional distribution of $Y|X$. The resulting sufficient summary plots can be particularly useful for guiding the choice of a first model at the beginning of analysis and for studying residuals after a model has been developed. In them, the predictor space over which Y is plotted 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 – *e.g.* convexity of the support of X. This intersection, called the central subspace $\mathcal{S}_{Y|X}$ for the regression of Y on X, becomes the natural focus of inferential interest, providing as it does the unique minimal sufficient summary plot over which the data can be viewed without any loss of regression information. Its dimension $d_{Y|X}$ is called the structural dimension of this regression. The book by Cook (1998b) provides a self-contained account and development of these foundational ideas.

Since the first moment-based methods, *sliced inverse regression* (Li 1991) and *sliced average variance estimation* (Cook and Weisberg 1991), were introduced, many others have been proposed. These 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 sliced inverse regression and sliced average

variance estimation, *Principal Hessian Directions* (Li, 1992; Cook, 1998a), *parametric inverse regression* (Bura and Cook 2001), 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 1998b). An exhaustiveness condition (recovery of the whole central subspace) is also required in some of these methods. *Average derivative estimation* (Härdle and Stoker 1989, Samarov 1993), the *structure adaptive method* (Hristache, Juditsky, Polzehl and Spokoiny 2001) and *minimum average variance estimation* (Xia, Tong, Li and Zhu 2002) 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) and *Fourier estimation* (Zhu and Zeng 2006), which may be flexibly regarded as inverse or forward, while requiring fewer assumptions.

Recently, Wang and Xia (2008) introduced *sliced regression*, reporting a range of advantages – estimation accuracy, exhaustiveness and robustness – over many other methods, albeit that it is limited to univariate Y and no sparse version is available. Whenever possible, we use this as the benchmark for the new approach presented here, the paper being organised as follows.

Section 2 introduces a joint approach combining speed with minimal assumptions. It targets the central subspace by exploiting a characterization of dimension reduction subspaces in terms of the Hellinger integral of order two (equivalently, χ^2-divergence), the immediacy and/or clarity of its local-through-global theory endorsing its naturalness (Sections 2.2 and 2.3). 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 (Section 2.2). Accordingly, our approach is more flexible than many others, multidimensional (discrete, continuous or mixed) Y, as well as X, being allowed (Section 2.1). Incorporating appropriate weights, it also unifies three existing methods, including sliced regression (Section 2.4).

Section 3 covers its implementation. 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 are in the Appendix.

Matlab code for all our algorithms are available upon request.

2 The Hellinger integral of order two

2.1 Terminology and notation

We establish here some basic terminology and notation.

We assume throughout that the $q \times 1$ response vector 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. Subject only to the central subspace existing and the finiteness condition on the Hellinger integral (below), both Y and X may be continuous, discrete or mixed. Where possible, we assay a unified account of such random vectors by phrasing development in terms of expectation and by adopting two convenient mild abuses of terminology. We use 'density' and the symbol $p(\cdot)$ to denote a probability density function, a probability function, or a mixture (product) of the two, and 'integral' – written $\int$ – to denote an integral in the usual sense, a summation, or a mixture of the two. Again, the argument of $p(\cdot)$ defines implicitly which distribution is being referred

to. Thus, in all cases,

$$p(w_1, w_2) = \ p(w_1|w_2)p(w_2) \quad (p(w_2) > 0)$$

where $p(w_1, w_2)$, $p(w_1|w_2)$ and $p(w_2)$ refer to the joint, conditional and marginal distributions of (W_1, W_2), $W_1|W_2$ and W_2 respectively.

The notation $W_1 \perp\!\!\!\perp W_2$ means that the random vectors W_1 and W_2 are independent. Similarly, $W_1 \perp\!\!\!\perp W_2|W_3$ 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 matrix representing orthogonal projection onto $\mathcal{S}$ with respect to the usual inner product while, for any x, $x_{\mathcal{S}}$ denotes its projection $P_{\mathcal{S}}x$. $\mathcal{S}(B)$, more fully $\mathrm{Span}(B)$, denotes the subspace of $\mathbb{R}^s$ spanned by the columns of the $s \times t$ matrix B. The trivial subspace comprising just the origin is thus denoted by $\mathcal{S}(0_s)$. 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 $\mathcal{A} \subseteq \mathcal{B}$ that $\mathcal{A}$ is a subset of $\mathcal{B}$, either $\mathcal{A} \subset \mathcal{B}$ or $\mathcal{A} = \mathcal{B}$, while $\|\cdot\|$ denotes the usual Euclidean norm for vectors or matrices.

2.2 Definition and four immediate properties

Throughout, u, u_1, u_2, ... denote fixed matrices with p rows. We develop here a population approach to dimension reduction in regression based on the Hellinger integral H of order two 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* $\frac{p(y,u^T x)}{p(y)p(u^T x)} = \frac{p(y|u^T x)}{p(y)} = \frac{p(u^T x|y)}{p(u^T x)}$ 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.

Four immediate properties of R and/or H show something of their potential utility here.

First, $H(u) = \int p(y|u^T x)p(u^T x|y)$ integrates information from forwards and inverse regression of Y on $u^T X$.

Second, 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, 1998b) is mirrored locally in $R(y^*; u^T x) = R(y; u^T x)$ and, hence, globally in $H(u; F_{(Y,X)}) = H(u; F_{(Y^*,X)})$.

Third, 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, 1998b) is mirrored locally in $R(y; u^T x) = R(y; (A^{-1}u)^T z)$ and, hence, globally in $H(u; F_{(Y,X)}) = H(A^{-1}u; F_{(Y,Z)})$. Because of these relations, there is no loss in standardizing the predictors to zero mean and identity covariance, which aids our algorithms.

Finally, there are clear links with departures from independence. 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, as is easily seen, the Hellinger integral is entirely equivalent to χ^2-divergence, large values of $H(u)$ reflecting strong dependence between Y and $u^T X$. Expectations being over the joint (Y, X) distribution, defining $\chi^2(u)$ as the common value of:

$$\int \left[\frac{\left\{ p(y, u^T x) - p(y)p(u^T x) \right\}^2}{p(y)p(u^T x)} \right] = \mathbb{E}\left[\frac{\left\{ R(Y; u^T X) - 1 \right\}^2}{R(Y; u^T X)} \right]$$

and noting that $\mathbb{E}\left[\left\{ R(Y; u^T X) \right\}^{-1} \right] = 1$, we have

$$\chi^2(u) = H(u) - 1. \tag{1}$$

Thus,

$$H(u) - 1 = \chi^2(u) \geq 0, \text{ equality holding if and only if } Y \perp\!\!\!\perp u^T X.$$

In particular, $H(0_p) = 1$.

2.3 Links with dimension reduction subspaces

The following results give additional properties confirming the Hellinger integral as a natural tool with which to study dimension reduction subspaces in general, and the central subspace $\mathcal{S}_{Y|X}$ in particular. The first establishes that $H(u)$ depends on u only via the subspace spanned by its columns.

Proposition 1 $\text{Span}(u_1) = \text{Span}(u_2) \Rightarrow R(y; u_1^T x) \overset{(y,x)}{\equiv} R(y; u_2^T x)$,
so that $H(u_1) = H(u_2)$ *and* $\chi^2(u_1) = \chi^2(u_2)$.

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, H and χ^2 themselves as with the following functions $\mathcal{R}_{(y,x)}$, $\mathcal{H}$ and $\mathcal{X}^2$ of a general subspace $\mathcal{S}$ which they induce. By Proposition 1, we may define

$$\mathcal{R}_{(y,x)}(\mathcal{S}) := R(y; u^T x),\ \mathcal{H}(\mathcal{S}) := H(u) \text{ and } \mathcal{X}^2(\mathcal{S}) := \chi^2(u)$$

where u is any matrix whose span is $\mathcal{S}$. Dependence properties at the end of the previous section give at once the following two results:

Proposition 2 $\mathcal{X}^2(\{0_p\}) = 0$. *That is,* $\mathcal{H}(\{0_p\}) = 1$.

Proposition 3 *For any subspace* $\mathcal{S}$ *of* $\mathbb{R}^p$,

$$\mathcal{H}(\mathcal{S}) - \mathcal{H}(\{0_p\}) = \mathcal{X}^2(\mathcal{S}) \geq 0,$$

equality holding if and only if $Y \perp\!\!\!\perp X_{\mathcal{S}}$.

Again, as 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 $1 \leq d \leq p$.

We seek now to generalise Proposition 3 from $(\{0_p\}, \mathcal{S})$ to any pair of nested subspaces $(\mathcal{S}_1, \mathcal{S}_1 \oplus \mathcal{S}_2)$, $\mathcal{S}_1$ and $\mathcal{S}_2$ meeting only at the origin. To do this, we introduce appropriate conditional quantities, as follows. Defining $\mathcal{X}^2(\mathcal{S}|\{0_p\})$ to be $\mathcal{X}^2(\mathcal{S})$, as is natural, Proposition 3 can be identified as the special case $(\mathcal{S}_1, \mathcal{S}_2) = (\{0_p\}, \mathcal{S})$ of:

$$\mathcal{H}(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{H}(\mathcal{S}_1) = \mathcal{X}^2(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{X}^2(\mathcal{S}_1) = \mathcal{X}^2(\mathcal{S}_2|\mathcal{S}_1) \geq 0,$$

equality holding if and only if $Y \perp\!\!\!\perp X_{\mathcal{S}_2}|X_{\mathcal{S}_1}$. Again, adopting the natural definition $\mathcal{X}^2(\{0_p\}|\mathcal{S}) := 0$, this same result holds at once (with equality) when $(\mathcal{S}_1, \mathcal{S}_2) = (\mathcal{S}, \{0_p\})$. Thus, it will suffice to establish it for nontrivial subspaces $(\mathcal{S}_1, \mathcal{S}_2)$.

Accordingly, let $\mathcal{S}_1 = \mathrm{Span}(u_1)$ and $\mathcal{S}_2 = \mathrm{Span}(u_2)$ 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 $\mathcal{R}_{(y,x)}$, $\mathcal{H}$ and $\mathcal{X}^2$, 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

$$R(y; u_1^T x, u_2^T x) = R(y; u_1^T x)R(y; u_2^T x|u_1^T x). \tag{2}$$

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

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

(2) gives:

$$H(u_1, u_2) = \mathbb{E}_{(Y,X)}\{R(Y; u_1^T X)H(u_2|u_1)\}, \tag{3}$$

while, noting 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 may put:

$$\mathcal{R}_{(y,x)}(\mathcal{S}_2|\mathcal{S}_1) := R(y; u_2^T x|u_1^T x) \text{ and, hence, } \mathcal{H}(\mathcal{S}_2|\mathcal{S}_1) := H(u_2|u_1).$$

Further, using Proposition 1 again, we may put $\mathcal{X}^2(\mathcal{S}_2|\mathcal{S}_1) := \chi^2(u_2|u_1)$ where $\chi^2(u_2|u_1)$ denotes the common value of

$$\begin{aligned}
&\int p(u_1^T x|y) \left[\frac{\left\{ p(y, u_2^T x|u_1^T x) - p(y|u_1^T x) p(u_2^T x|u_1^T x) \right\}^2}{p(y|u_1^T x) p(u_2^T x|u_1^T x)} \right] \\
&= \mathbb{E}_{(Y,X)} R(Y; u_1^T X) \left[\frac{\left\{ R(Y; u_2^T X|u_1^T X) - 1 \right\}^2}{R(Y; u_2^T X|u_1^T X)} \right].
\end{aligned}$$

Noting that $\mathbb{E}_{u_2^T X|(Y,u_1^T X)} \left[\left\{ R(Y; u_2^T X|u_1^T X) \right\}^{-1} \right] = 1$, and recalling (3), it follows that (*cf.* (1)):

$$\chi^2(u_2|u_1) = \mathbb{E}_{(Y,X)} R(Y; u_1^T X) \{ H(u_2|u_1) - 1 \} = H(u_1, u_2) - H(u_1).$$

Recalling Propositions (1) to (3), we have shown, as desired:

Proposition 4 *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces of $\mathbb{R}^p$ meeting only at the origin. Then,*

$$\mathcal{H}(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{H}(\mathcal{S}_1) = \mathcal{X}^2(\mathcal{S}_1 \oplus \mathcal{S}_2) - \mathcal{X}^2(\mathcal{S}_1) = \mathcal{X}^2(\mathcal{S}_2|\mathcal{S}_1) \geq 0,$$

equality holding 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.

Moreover, they can be used to characterize dimension reduction subspaces and, thereby, the central subspace $\mathcal{S}_{Y|X} = \text{Span}(\eta)$ say, where η has full column rank $d_{Y|X} := \dim(\mathcal{S}_{Y|X})$, the structural dimension of the regression of Y on X.

Theorem 5 *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 5 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 $\mathcal{U}_d$ 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. Whereas, for $d > 0$, η_d is at best unique up to post-multiplication by an orthogonal matrix, $\text{Span}(\eta_d)$ *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 5 give at once:

Corollary 6 *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. $d = 0 \Rightarrow \left[\overline{H}_d = 1 \text{ and } \mathbb{S}_d = \{\{0_p\}\} \right]$.

We also have:

Proposition 7 $d_1 < d_2 \leq d_{Y|X} \Rightarrow 1 \leq \overline{H}_{d_1} < \overline{H}_{d_2}$.

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. The next section discusses such estimates, and algorithms to compute them. Before that, we note that ours is a unifying approach.

2.4 A unifying approach

Our approach unifies three other methods, in the sense that each can be shown to be equivalent to adopting suitably weighted forms of the order two Hellinger integral. Details are given in the second part of the Appendix.

The methods are: (a) for sliced Y, kernel discriminant analysis as developed by Hernánedez and Velilla (2005), whose inverse method uses a global search; (b) sliced regression (Wang and Xia, 2008), whose forwards method uses a local search; and (c) density minimum average variance estimation (Xia, 2007), which again uses a local approximation.

Of these, our local approach, developed next, is much faster than kernel discriminant analysis, which also has the disadvantage of depending heavily on starting values. Again,

Wang and Xia (2008) argue for sliced regression in preference to density minimum average variance estimation, this latter method assuming $d_{Y|X}$ known, requiring Y to be continuous, and generally being found to be slower and less accurate. Accordingly, it will suffice to compare our new approach with sliced regression, whenever the latter can be used.

3 Estimation procedure

Having defined our weighted approximation (Section 3.1), we first assume that $d_{Y|X}$ is known and propose an estimation procedure for $\mathcal{S}_{Y|X}$ (Section 3.2). We then suggest a permutation test to estimate $d_{Y|X}$ (Section 3.3). We also propose a sparse version of the method, particularly useful when only a few of the predictor variables are related to the response (Section 3.4). We will use Y as a scalar for general discussion, but discrete and vector Y are treated as required. As noted in Section 2.2, X can be taken as standardized without loss.

Although similar in spirit to Xia (2007) and Wang and Xia (2008), rather than localize X and slice Y, the approach taken here localizes (X, Y) in the sense that we directly approximate the Hellinger integral via a local approach. This brings a number of benefits, including greater speed, robustness (see, in particular, Example 1 below) and better handling of cases where Y takes only a few discrete values, (see point (b), at the end of the following section).

In practice, real data sets can be subject to a considerable amount of inherent randomness or noise. A further advantage of our approach is that it is straightforward to adapt it to gain efficiency in such contexts by a form of 'thresholding', as described in Section 4.8 (Example 8).

3.1 Weighted approximation

To use the Hellinger index, we have to estimate $\frac{p(x,y)}{p(x)p(y)}$. Hence, estimation of $p(\cdot)$ is critical. For convenience of derivation and without loss of generality, suppose that $d_{Y|X} = 1$ and consider a particular point (x_0^*, y_0^*):

$$\frac{p(x_0^*, y_0^*)}{p(x_0^*)p(y_0^*)} = \frac{p(\eta^T x_0^*, y_0^*)}{p(\eta^T x_0^*)p(y_0^*)}.$$

We put $(x_0, y_0) := (\eta^T x_0^*, y_0^*)$, denoting a corresponding general point by (x, y). Let $w_0(x, y) := \frac{1}{h_1}K(\frac{x-x_0}{h_1})\frac{1}{h_2}K(\frac{y-y_0}{h_2})$, where $K(.)$ is a smooth kernel function, symmetric about 0, h_1 and h_2 being corresponding bandwidths. Define $s_2 := \int u^2 K(u)du$ and $g_{ij} := \int u^i v^j K(u)K(v)p(x_0 - h_1 u, y_0 - h_2 v)dudv$, so that

$$g_{00} \sim p(x_0, y_0), g_{10} \sim -h_1 s_2 p^{10}(x_0, y_0), g_{01} \sim -h_2 s_2 p^{01}(x_0, y_0),$$

$$g_{20} \sim s_2 p(x_0, y_0), g_{02} \sim s_2 p(x_0, y_0), g_{11} \sim h_1 h_2 s_2^2 p^{11}(x_0, y_0),$$

$$g_{22} \sim s_2^2 p(x_0, y_0), g_{21} \sim -h_2 s_2^2 p^{01}(x_0, y_0), g_{12} \sim -h_1 s_2^2 p^{10}(x_0, y_0),$$

where p^{ij} are the corresponding derivatives for the density. Then,

$$\begin{aligned}
\mathbb{E}w_0(x, y) &= p(x_0, y_0) \text{ and} \\
\mathbb{E}w_0(x, y)xy &= \int p(x, y)\frac{1}{h_1}K(\frac{x - x_0}{h_1})\frac{1}{h_2}K(\frac{y - y_0}{h_2})xydxdy \\
&= \int p(x_0 - h_1 u, y_0 - h_2 v)K(u)K(v)(x_0 - h_1 u)(y_0 - h_2 v)dudv \\
&= x_0 y_0 g_{00} - h_1 y_0 g_{10} - x_0 h_2 g_{01} + h_1 h_2 g_{11},
\end{aligned}$$

while

$$\begin{aligned}
\mathbb{E}w_0(x, y)x^2y^2 &= \int p(x, y)\frac{1}{h_1}K(\frac{x - x_0}{h_1})\frac{1}{h_2}K(\frac{y - y_0}{h_2})x^2y^2dxdy \\
&= \int p(x_0 - h_1 u, y_0 - h_2 v)K(u)K(v)(x_0 - h_1 u)^2(y_0 - h_2 v)^2dudv \\
&= x_0^2 y_0^2 g_{00} - 2h_1 x_0 y_0^2 g_{10} + h_1^2 y_0^2 g_{20} - 2x_0^2 h_2 y_0 g_{01} + 4h_1 h_2 x_0 y_0 g_{11} \\
&\quad - 2h_2 y_0 h_1^2 g_{21} + x_0^2 h_2^2 g_{02} - 2h_1 x_0 h_2^2 g_{12} + h_1^2 h_2^2 g_{22}.
\end{aligned}$$

Similarly, with $w_0(x) = \frac{1}{h_1}K(\frac{x-x_0}{h_1})$,

$$\mathbb{E}w_0(x) = p(x_0) \text{ and}$$

$$\mathbb{E}w_0(x)x = x_0p(x_0) + h_1^2 s_2 p'(x_0),$$

while

$$\mathbb{E}w_0(x)x^2 = x_0^2p(x_0) + 2h_1^2 s_2 x_0 p'(x_0) + h_1^2 s_2 p(x_0).$$

Thus,

$$\frac{p(x_0, y_0)}{p(x_0)p(y_0)} \sim \frac{h_1^2 h_2^2 s_2^2}{h_1^2 h_2^2 s_2^2 + h_1^2 y_0^2 s_2 + h_2^2 x_0^2 s_2} H^*$$

where

$$H^* := \frac{\mathbb{E}w_0(x,y)x^2y^2 - (\mathbb{E}w_0(x,y)xy)^2/\mathbb{E}w_0(x,y)}{[\mathbb{E}w_0(x)x^2 - (\mathbb{E}w_0(x)x)^2/\mathbb{E}w_0(x)][\mathbb{E}w_0(y)y^2 - (\mathbb{E}w_0(y)y)^2/\mathbb{E}w_0(y)]}.$$

Again, without loss of generality, we can assume $(x_0, y_0) = (0, 0)$ (otherwise, set $x' = x - x_0$ and $y' = y - y_0$, and use the argument with (x', y') instead) so that, finally,

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

Pooling first, we may find η to maximize $\sum_{i=1}^n H_i^*$, where H_i^* is the H^* value for the i^{th} data point. Alternatively, we may maximize H_i^* to obtain η_i at each data point, and then pool the η_i's together. Here, we adopt the latter approach.

Clearly, the method will depend on the choice of w_0. It is interesting to note that, if all $w_0(x, y) \equiv 1$, η is the dominant eigenvector of $V(X)^{-1}V(XY)V(Y)^{-1}$, where $V(\cdot)$ denotes a variance matrix. Suppose that both X and Y are standardized, then $V(X)^{-1}V(XY)V(Y)^{-1} = \mathbb{E}\{XX^TV(Y|X)\} + V[X\mathbb{E}(Y|X)]$, which may be regarded as weighted from the regression mean and variance functions. Further, assume that the linearity and constant variance conditions hold. Then, as an inverse regression,

$$\begin{aligned} V(X)^{-1}V(XY)V(Y)^{-1} = &\quad \mathbb{E}\{Y^2V(X|Y)\} + V[Y\mathbb{E}(X|Y)] \\ = &\quad \mathbb{E}(Y^2)(I - P_{\mathcal{S}_{Y|X}}) + P_{\mathcal{S}_{Y|X}}V(XY)P_{\mathcal{S}_{Y|X}} \end{aligned}$$

which may be simply regarded as the weighted inverse mean and variance functions. However, the corresponding η may not always be in the central subspace $\mathcal{S}_{Y|X}$, underlining the importance of the choice of weights.

In this paper, we explore the approach obtained by choosing w_0 to be 1 for a case within a Euclidean k-nearest-neighborhood (kNN) and 0 else, as follows:

- (a) *Continuous – or discrete, but meaningful, numerical – 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 k nearest-neighbors of (X_i, Y_i).

- (b) *Categorical response $Y \in \{1, ..., C\}$:* In this case, we find the best directions to discriminate between the local conditional predictor variances, using:

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

  where, here, a subscript 'ki' denotes computation over the k nearest-neighbors of X_i, *discarding* this term from the analysis if none of the k corresponding Y values differs from Y_i. Not having this condition – thereby, losing discriminatory power (effectively, being distracted by noise) – and using a different choice of localization, sliced regression tends to perform less well here, especially when the number of categories C is small.

- (c) *Multivariate response Y:* In this case, rather than adopt the projective resampling idea of Li, Wen and Zhu (2008), we use the following, fast approach, acting as if the elements of $(Y|X = x)$ were independent. A subscript 'ki' denoting computation over the k nearest-neighbors of (X_i, Y_i), as in (a), the above local univariate calculation generalises at once to:

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

in which the scalar factor involving $V_{ki}(Y^T) := \mathbb{E}_{ki}(Y^T Y) - \mathbb{E}_{ki}(Y^T)\mathbb{E}_{ki}(Y)$ is ignorable as only orthonormal eigenvectors of $H_i^*(k)$ are used, as explained in the following section.

3.2 Algorithm

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, the algorithm is as follows:

1. For each observation (x_i, y_i), find its k nearest neighbors in terms of the Euclidean distance $\|(x, y) - (x_i, y_i)\|$ ($\|x - x_i\|$, if Y is categorical) and, hence, the $p \times d$ orthonormal matrix $\eta_i \equiv (\eta_{i1}, \eta_{i2}, \cdots, \eta_{id})$ whose columns contain the d dominant eigenvectors of $H_i^*(k)$.

2. Find the spectral decomposition of $\hat{M} := \frac{1}{nd}\sum_{i=1}^{n} \eta_i \eta_i^T$, 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 this paper, a rough choice of k around $2p$ seems to work well across the range of examples presented below. More refined ways to choose k, such as cross-validation, could of course be used at greater computational expense.

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

Recall that $d_{Y|X} = 0$ is equivalent to $Y \perp\!\!\!\perp X$. In the population, the orthogonal matrix of 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, its eigenvalues reflecting

the magnitude of dependence between Y and the corresponding regressors $\beta^T X$. In the sample, holding fixed the observed responses $\mathbf{y} := (y_1, ..., y_n)^T$ while randomly permuting the rows of the $n \times p$ matrix $\mathbf{X} := (x_1, ..., x_n)^T$ will change $\hat{M}$, tending to reduce the magnitude of the corresponding observed dependencies – except, that is, 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 $m \in \{0, \cdots, p-1\}$. Sampling variability in (B_m, A_m) apart, this is equivalent to testing $Y \perp\!\!\!\perp A_m^T X | B_m^T X$, where $B_m := (\hat{\beta}_1, \cdots, \hat{\beta}_m)$ and $A_m := (\hat{\beta}_{m+1}, \cdots, \hat{\beta}_p)$. Accordingly, the following procedure can be used to determine $d_{Y|X}$:

- Obtain $\hat{M}$ from the original $n \times (p+1)$ data matrix $(\mathbf{X}, \mathbf{y})$, computing its spectrum $\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots > \hat{\lambda}_p$ and, from it, 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, obtaining from each a new matrix $\hat{M}_j$ and, hence, a new test statistic f_j, $(j = 1, \cdots, J)$.

- Compute the permutation p-value:

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

  rejecting H_0 if $p_{perm} < \alpha$.

- Repeat the last two steps for $m = 0, 1, \ldots$ until H_0 cannot be rejected. 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. In such cases, effectively

selecting informative predictors in the reduced directions can improve both estimation accuracy and interpretability. In this section, we incorporate the shrinkage estimation procedure proposed by Li and Yin (2008) into our method, assuming again that $d_{Y|X} = d$ is known.

Our standard algorithm uses the estimate $\widehat{\mathcal{S}}_{Y|X} = \text{Span}(\hat{u})$, $\hat{u} := (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_d)$ containing the dominant d eigenvectors of

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

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

$$\tilde{u} := \underset{u \in \mathcal{U}_d}{\arg\min}\, g(u) \text{ where } g(u) := \sum\nolimits_{i=1}^{n} \left\| \eta_i - u\, u^T \eta_i \right\|^2, \tag{4}$$

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

To select informative predictors, a shrinkage index vector α can be incorporated into this alternative formulation (4), as follows. With $\alpha \in R^p$ constrained by $\sum_{i=1}^{p} |\alpha_i| \leq \lambda$ for some $\lambda > 0$, let $\hat{\alpha}$ minimise

$$\sum_{i=1}^{n} ||\eta_i - diag(\alpha) \hat{u}\, \hat{u}^T \eta_i||^2, \tag{5}$$

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

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

where RSS_λ is the residual sum of squares from the fit in (5), p_λ being the number of non-zero elements in $\hat{\alpha}$.

4 Evaluation

We evaluate the performance of our method on both simulated and real data.

To measure the accuracy with which Span($\hat{u}$) estimates Span(u), we use each of three measures proposed in the literature: the vector correlation coefficient, $r = \left|\det(u^T\hat{u})\right|$ (Ye and Weiss, 2003), the Euclidean distance between projectors, $\Delta(u,\hat{u}) = \left\|uu^T - \hat{u}\hat{u}^T\right\|$ (Li, Zha and Chiaromonte, 2005), and the residual length $m^2 = \left\|(I-uu^T)\hat{u}\right\|$ (Xia et al, 2002), this last being decomposable by dimension:

$$m_r^2 = \left\|(I-uu^T)\hat{u}_r\right\| \text{ where } \hat{u} \equiv (\hat{u}_1, ..., \hat{u}_d).$$

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. All simulation results are based on 100 replications.

To evaluate the effectiveness of our approach, we compare our results with sliced regression (SR) whenever this method is available, it being reported by Wang and Xia (2008) to have many advantages (estimation accuracy, exhaustiveness and robustness) in finite sample performances over many other methods.

4.1 Example 1: A model with frequent extremes

Following Wang and Xia (2008, example 1), $Y = (X^T\beta)^{-1} + 0.2\epsilon$, where $X = (x_1, ..., x_{10})^T$ and ϵ i.i.d. standard normal random variables, while $\beta = (1,1,1,1,0,0,0,0,0,0)^T$, so that $S_{Y|X} = \text{span}(\beta)$. Samples of size $n = 200, 400$ and 800 are used.

Note that extreme Y occurs often in this model, (in particular, for covariate and error values around their means). Because of this, some existing dimension reduction methods – such as minimum average variance estimation, sliced inverse regression and sliced average variance estimation – don't perform well here, sliced regression standing out as the best method in Wang and Xia (2008).

Its localization of (X, Y) filtering the extremes, our method is very robust here. Applying it to this model, we obtain the results detailed in Table 1.

Table 1: Accuracy of estimates

n	r	$\Delta(u, \hat{u})$	m^2	CPU time (in seconds)
200	0.991(0.007)	0.126(0.041)	0.020(0.011)	8
(SR)	0.998(0.001)	0.062(0.022)	0.004(0.002)	265
400	0.996(0.002)	0.079(0.018)	0.007(0.004)	20
(SR)	0.999(0.001)	0.041(0.013)	0.002(0.001)	789
800	0.999(0.001)	0.049(0.012)	0.002(0.001)	57
(SR)	0.999(0.001)	0.026(0.005)	0.001(0.001)	2,519

Overall, estimation accuracy is better for sliced regression, while our method shows a very significant computational gain. Both the relative accuracy and the relative speed of our method increase with sample size.

4.2 Example 2: A sparse model

Following Wang and Xia (2008, example 2), who in turn follow Li (1992, model (8.1)), $Y = \cos(2X_1) - \cos(X_2) + 0.2\epsilon$, where $X_1, ..., X_{10}$ and ϵ are i.i.d. standard normal random variables, so that $S_{Y|X} = \text{span}(e_1, e_2)$. Samples of size $n = 200, 400$ and 800 are used.

As in the previous example, Table 2 shows that, at very significant computational cost, the estimation accuracy of sliced regression is greater than for our standard

method, whose relative accuracy and speed increase with sample size. Overall, our sparse method is most accurate, as might be expected. While always good, its TPR and FPR (reported in Table 3) improve with n.

Table 2: Accuracy of estimates

n	r	$\Delta(u, \hat{u})$	m_1^2	m_2^2	CPU time (in seconds)
200	0.772(0.162)	0.564(0.178)	0.060(0.031)	0.336(0.213)	8
(Sparse)	0.942(0.183)	0.161(0.231)	0.004(0.012)	0.076(0.198)	
(SR)	0.921(0.112)	0.313(0.171)	0.011(0.007)	0.172(0.013)	269
400	0.894(0.105)	0.456(0.151)	0.021(0.010)	0.189(0.137)	20
(Sparse)	0.981(0.081)	0.092(0.146)	0.001(0.002)	0.029(0.103)	
(SR)	0.988(0.038)	0.131(0.054)	0.002(0.001)	0.022(0.009)	825
800	0.945(0.029)	0.301(0.078)	0.009(0.005)	0.073(0.044)	58
(Sparse)	0.993(0.014)	0.069(0.093)	0.0004(0.001)	0.013(0.027)	
(SR)	0.992(0.010)	0.094(0.023)	0.001(0.001)	0.009(0.005)	2,700

Table 3: Effectiveness of variable selection

	Example 2		Example 5	
n	TPR	FPR	TPR	FPR
200	0.955	0.331	0.938	0.400
400	0.995	0.210	0.974	0.300
800	1.000	0.115	1.000	0.194

4.3 Example 3: A categorical response model

Following Zhu and Zeng (2006), $Y = I[X\beta_1 + 0.2\epsilon > 1] + 2I[X\beta_2 + 0.2\epsilon > 0]$, where $X = (x_1, ..., x_{10})$ and ϵ are i.i.d. standard normal random variables, while $\beta_1 = (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^T$ and $\beta_2 = (0, 0, 0, 0, 0, 0, 1, 1, 1, 1)^T$, so that $S_{Y|X} = \text{span}(\beta_1, \beta_2)$. Samples of size $n = 200, 400$ and 800 are used.

This model has a discrete response y, which we treat as categorical, taking values in $\{0, 1, 2, 3\}$. Thus, sliced regression can take at most $C = 4$ slices. In our experience, small values of C can affect the performance of sliced regression dramatically, (*cf.* (b) of Section 3.1 above). Table 4 shows that, overall, our method is more accurate, even for small samples – and this while maintaining its significant speed advantage, (which again grows with n). See also Example 4 below.

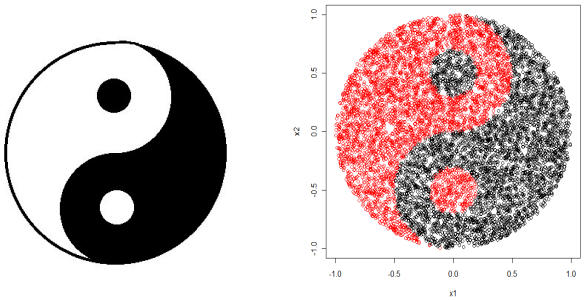
Table 4: Accuracy of estimates

n	r	$\Delta(u, \hat{u})$	m_1^2	m_2^2	CPU time (in seconds)
200	0.953(0.017)	0.257(0.049)	0.043(0.021)	0.051(0.025)	24
(SR)	0.933(0.051)	0.308(0.114)	0.022(0.012)	0.101(0.088)	265
400	0.978(0.009)	0.175(0.039)	0.020(0.012)	0.024(0.012)	53
(SR)	0.976(0.013)	0.187(0.061)	0.008(0.004)	0.037(0.026)	809
800	0.991(0.003)	0.115(0.021)	0.009(0.005)	0.009(0.005)	128
(SR)	0.991(0.004)	0.110(0.031)	0.003(0.002)	0.012(0.007)	2,607

4.4 Example 4: A binary response model

Consider the well-known *Tai Chi* figure in Asian culture shown in the lefthand 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.

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 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.



(a) Tai Chi (b) Projection onto x1 and x2

Figure 1: Tai Chi figure

Table 5: Accuracy of estimates: Tai Chi data

n	r	$\Delta(u, \hat{u})$	m_1^2	m_2^2
200	0.995(0.003)	0.084(0.023)	0.006(0.004)	0.004(0.003)
(SR)	0.409(0.231)	0.965(0.142)	0.714(0.206)	0.023(0.011)
400	0.998(0.001)	0.056(0.014)	0.003(0.002)	0.002(0.001)
(SR)	0.440(0.269)	0.837(0.174)	0.713(0.254)	0.009(0.005)
800	0.999(0.0003)	0.037(0.007)	0.001(0.001)	0.001(0.001)
(SR)	0.514(0.305)	0.766(0.234)	0.633(0.306)	0.004(0.002)

Li (2000) analyzed this example from the perspective of dimension reduction. Due to the binary response, sliced inverse regression can only find 1 direction and so he proposed a double slicing scheme to identify the second direction. Here, we apply our method and sliced regression to these data. Both our permutation test and cross-validation indicate a structural dimension of two. Table 5 shows that our much faster method is also much more accurate here, sliced regression effectively missing the second direction (reasons for its relatively poor performance in contexts such as this having been discussed above).

4.5 Example 5: High dimensional instances of a sparse model

Following Xia (2007, example 4.2), $Y = 2(X^T\beta_1)+2\exp(X^T\beta_2)\epsilon$, where $X = (x_1, ..., x_{10})^T$, $\{x_r\}_{r=1}^{10} \sim \text{Uniform}(-\sqrt{3}, \sqrt{3})$ and $\epsilon \sim N(0,1)$ are independent, while

$$\beta_1 = (1, 2, 0, \cdots, 0, 2)^T/3 \text{ and } \beta_2 = (0, 0, 3, 4, 0, \cdots, 0)^T/5,$$

so that $S_{Y|X} = \text{span}(\beta_1, \beta_2)$. Note that, in this example, the predictors are uniformly distributed, while one direction in the central subspace is in the variance function. High dimensional instances are included.

Table 6 shows that sliced regression is more accurate, when it is available, but at an even higher computational cost than in previous examples. Indeed, it stopped altogether when $n = 1000$ and $p = 50$. Our method continues to perform well for high-dimensional problems, the sparse version delivering improved performance, as anticipated. Its TPR and FPR are reported in Table 3.

Table 6: Accuracy of estimates

n	p	r	$\Delta(u, \hat{u})$	m_1^2	m_2^2	CPU time (in seconds)
400	10	0.885(0.066)	0.399(0.106)	0.076(0.047)	0.145(0.104)	22
(Sparse)		0.915(0.044)	0.351(0.111)	0.056(0.044)	0.109(0.077)	
(SR)		0.961(0.012)	0.232(0.063)	0.017(0.018)	0.055(0.028)	845
800	10	0.949(0.018)	0.270(0.056)	0.039(0.019)	0.062(0.032)	62
(Sparse)		0.963(0.021)	0.232(0.069)	0.027(0.019)	0.046(0.035)	
(SR)		0.981(0.009)	0.161(0.041)	0.010(0.006)	0.024(0.013)	2,713
	20	0.907(0.030)	0.353(0.059)	0.064(0.023)	0.119(0.049)	97
(Sparse)		0.953(0.031)	0.255(0.083)	0.044(0.034)	0.049(0.044)	
(SR)		0.955(0.012)	0.238(0.042)	0.022(0.007)	0.057(0.022)	11,159
	50	0.725(0.104)	0.590(0.109)	0.176(0.051)	0.347(0.146)	409
(Sparse)		0.876(0.184)	0.364(0.191)	0.069(0.046)	0.122(0.171)	
(SR)		0.867(0.031)	0.431(0.061)	0.063(0.017)	0.186(0.051)	43,021
1000	50	0.815(0.045)	0.492(0.068)	0.129(0.037)	0.235(0.069)	570
(Sparse)		0.955(0.031)	0.249(0.081)	0.049(0.038)	0.041(0.036)	
(SR)		NA				
1500	50	0.891(0.023)	0.372(0.044)	0.083(0.019)	0.129(0.029)	1,243
(Sparse)		0.973(0.018)	0.190(0.059)	0.028(0.024)	0.025(0.020)	
(SR)		NA				

4.6 Example 6: Determining structural dimension

Here, we report the finite-sample performance of our approach to determining structural dimension (Section 3.3).

The results in Table 7 are based on 100 data sets with sample size 400, as used in the previous examples. The significance level is $\alpha = 0.05$, while $J = 1000$ permutations are used. The numbers in boldface are the percentages of correctly identified $d_{Y|X}$, which seem very reasonable.

Table 7: Permutation test for d

	Percentage of estimated dimension			
Example	$d=0$	$d=1$	$d=2$	$d=3$
1	0	**0.97**	0.03	
2	0	0.14	**0.85**	0.01
3	0	0.12	**0.88**	0.00
4	0	0.21	**0.70**	0.09

4.7 Example 7: A multivariate response model

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

$$Y_1 = 1/(X^T\beta_1) + 0.5\epsilon_1, Y_2 = 2\exp(X^T\beta_2)\epsilon_2, Y_3 = \epsilon_3 \text{ and } Y_4 = \epsilon_4,$$

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

The sliced regression method being unavailable here, we compare our (standard) method, our method together with the projective re-sampling (PR-H2) approach of Li, Wen and Zhu (2008), and projective resampling with sliced inverse regression (PR-SIR). Following Li, Wen and Zhu (2008), the numbers of slices used are 10, 20 and 20 corresponding, respectively, to sample sizes of 200, 400 and 800. The Monte Carlo sample size is m_n=2000 for the PR type approaches, results being given in Table 8.

Overall, our standard method is faster and more accurate than PR-SIR. Projective re-sampling does not improve its performance, despite taking longer. As expected, the estimation accuracy of all methods improves with sample size.

Table 8: Accuracy of estimates

n	r	$\Delta(u, \hat{u})$	m_1^2	m_2^2
200	0.766(0.186)	0.545(0.201)	0.109(0.077)	0.297(0.242)
(PR-H2)	0.790(0.203)	0.506(0.208)	0.089(0.058)	0.265(0.255)
(PR-SIR)	0.746(0.233)	0.605(0.187)	0.050(0.032)	0.401(0.249)
400	0.964(0.018)	0.225(0.057)	0.033(0.020)	0.039(0.029)
(PR-H2)	0.961(0.021)	0.233(0.063)	0.036(0.019)	0.042(0.032)
(PR-SIR)	0.894(0.095)	0.418(0.127)	0.022(0.011)	0.191(0.125)
800	0.988(0.005)	0.133(0.028)	0.012(0.007)	0.013(0.007)
(PR-H2)	0.988(0.004)	0.132(0.025)	0.011(0.004)	0.013(0.007)
(PR-SIR)	0.961(0.037)	0.253(0.068)	0.011(0.006)	0.067(0.038)

4.8 Example 8: 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 Table 9.

Table 9: 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	-0.01	0.18	-0.33
x_2	median family income	0.19	0.14	-0.43
x_3	percentage of people under the poverty level	0.92	-0.30	0.07
x_4	unemployment rate	0.04	-0.08	-0.10
x_5	percentage of population who are divorced	0.07	0.23	-0.51
x_6	percentage of kids born to never married	-0.27	-0.82	-0.41
x_7	percentage of people who speak only English	0.02	0.04	-0.00
x_8	mean persons per household	0.06	0.21	-0.32
x_9	percentage of people in owner occupied households	-0.01	-0.18	-0.22
x_{10}	percentage of housing occupied	-0.00	-0.19	0.18
x_{11}	median value of owner occupied house	-0.09	-0.10	0.26
x_{12}	population density in persons per square mile	0.13	-0.02	-0.11
x_{13}	percent of people using public transit for commuting	-0.01	0.04	0.01

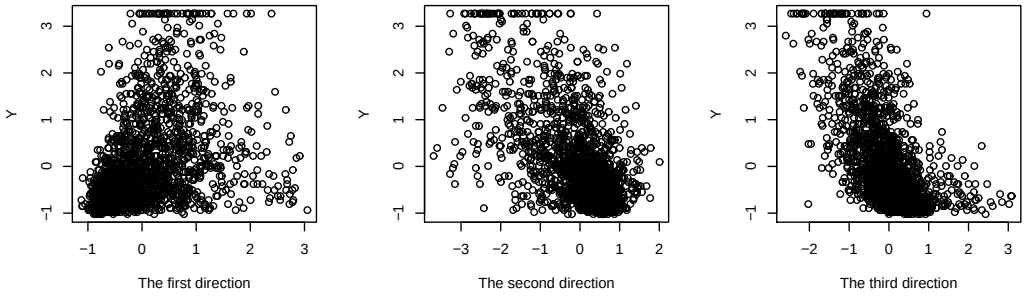


Figure 2: Community and crime

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 the sliced regression.

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 sum of the largest d eigenvalues in η_i exceeds a specified threshold. Here, we use 0.65, 0.75 and 0.85 as threshold values for d=1, 2 and 3 respectively, the permutation test giving p-values of 0, 0, 0.022 and 0.132 for d=0, 1, 2 and 3.

With d=3, we find the direction estimates reported in Table 9. The first direction is dominated by x_3, the percentage of people under the poverty level, and the second by 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. Among other family structure variables, the divorce rate is another important factor contributing to community crime.

5 Discussion

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

Overall, 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 weight approximation. Relative to sliced regression (Wang and Xia, 2008), examples show that our approach: (a) is significantly faster without losing too much estimation accuracy, allowing larger problems to be tackled, (b) is more general, multidimensional (discrete, continuous or mixed) Y as well as X 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.

Among other further work, a global search of the Hellinger integral with or without slicing, similar to that of Hernánedez and Velilla (2005), merits investigation. That said, being based on nonparametric density estimation, this may turn out to be slow, depend heavily on starting values, or be less accurate than a local approach, in particular when $d_{Y|X} > 1$.

6 Appendix: Additional materials

In this section, we provide additional materials. First part is the proof of the Theoretical results. Second part is the links to some existing methods.

6.1 Additional justifications

We provide here justifications of results not given in the body of the paper.

Proposition 1

It suffices to show the first implication. Now, $\text{Span}(u_1) = \text{Span}(u_2) \Rightarrow \text{rank}(u_1) = \text{rank}(u_2) = r$, say. Suppose first that $r = 0$. Then, for $i = 1, 2$, u_i vanishes, so that $Y \perp\!\!\!\perp u_i^T X$ implying $R(y; u_i^T x) \overset{(y,x)}{=} 1$. 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 $R(y; u_1^T x) \overset{(y,x)}{\equiv} R(y; u_2^T x)$.

Theorem 5

Since the central subspace is the intersection of all dimension reduction subspaces, it suffices to prove (1). If $\mathcal{S} = \mathbb{R}^p$, the result is trivial. Again, if $\mathcal{S} = \{0_p\}$, it follows at once from Proposition 3. Otherwise, it follows from Proposition 4, taking $\mathcal{S}_2$ as the orthogonal complement in $\mathbb{R}^p$ of $\mathcal{S}_1 = \mathcal{S}$.

Proposition 7

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

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

the second inequality here using Proposition 4. Suppose, if possible, that $\overline{H}_{d_1} = \overline{H}_{d_2}$. Then, $H(\eta_{d_1}, u) = H(\eta_{d_1})$ for any such u so that, using Proposition 4, $Y \perp\!\!\!\perp v^T X | \eta_{d_1}^T X$ for any vector v. It follows that $\text{Span}(\eta_{d_1})$ is a dimension reduction subspace, contrary to $d_1 < d_{Y|X}$. The proof when $d_1 = 0$ is entirely similar, using Proposition 3.

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

$$p(u^T x) = \sum_{y \in \mathcal{Y}} p(y, u^T x) \text{ where } p(y, u^T x) = p(y)p(u^T x|y) = p(u^T x)p(y|u^T x). \quad (6)$$

In the discrete case, it is natural to use the following form of $H(\cdot)$ as our basis for estimation:

$$H(u) = \mathbb{E}\left(\frac{p(u^T X|Y)}{p(u^T X)}\right).$$

For example, in the binary case, by (6), we need just two m-dimensional estimates:

$$\widehat{p}(u^T x|Y = 0) \text{ and } \widehat{p}(u^T x|Y = 1)$$

obtained from (kernel) smoothing the corresponding partition of the data.

Thus,

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

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,

$$\mathbb{E}_{(X,Y)}\left(\frac{p(Y|X)}{p(Y)}\right) = \mathbb{E}_X \sum_{i=1}^{k}\left(\frac{(p(C_i|X))^2}{p(C_i)}\right) = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}_X\left\{\mathbb{E}^2_{Y|X}(I_{C_i}(Y)|X)\right\} \quad (7)$$

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

$$k = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}(I^2_{C_i}(Y)) = \sum_{i=1}^{k}\frac{1}{p(C_i)}\mathbb{E}_X\mathbb{E}_{Y|X}(I^2_{C_i}(Y)|X). \quad (8)$$

Denoting the sliced form of Y by $\tilde{Y}$, (7) and (8) 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, optimising the Hellinger integral of order 2 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)

Note that, 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, density minimum average variance estimation and density outer product of gradient (Xia, 2007) are methods to estimate the last term, apart from the weight $p(y)^{-2}$.

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.

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

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

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 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 & 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.

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.

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.

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. (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., 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.

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.