Electronic Journal of Statistics
Vol. 5 (2011) 1471–1494
ISSN: 1935-7524
DOI: 10.1214/11-EJS650

Iterative application of dimension reduction methods

Amanda J. Shaker and Luke A. Prendergast

Department of Mathematics and Statistics
La Trobe University
Bundoora, VIC 3086, Australia
e-mail: ajshaker@students.latrobe.edu.au; luke.prendergast@latrobe.edu.au

Abstract: The goal of this article is to introduce an iterative application of dimension reduction methods. It is known that in some situations, methods such as Sliced Inverse Regression (SIR), Ordinary Least Squares (OLS) and Cumulative Mean Estimation (CUME) are able to find only a partial basis for the dimension reduction subspace. However, for many models these methods are very good estimators of this partial basis. In this paper we propose a simple iterative procedure which differs from existing combined approaches in the sense that the initial partial basis is estimated first and the second dimension reduction approach seeks only the remainder of the dimension reduction subspace. Our approach is compared against that of existing combined dimension reduction approaches via simulated data as well as two example data sets.

AMS 2000 subject classifications: Primary 62J02; secondary 62H30.
Keywords and phrases: Dimension reduction, sliced inverse regression, sliced average variance estimation, ordinary least squares, principal Hessian directions, cumulative mean estimation, cumulative variance estimation, cumulative directional regression.

Received May 2011.

Contents

1 Introduction 1472
2 Dimension reduction 1473
2.1 Ordinary least squares 1473
2.2 Sliced inverse regression 1474
2.3 Sliced average variance estimation 1475
2.4 Principal Hessian directions 1475
2.5 Combined approaches 1476
2.6 Cumulative slicing procedures 1476
3 Iterative use of inverse regression methods 1477
3.1 A motivating example 1477
3.2 Theory 1478
3.3 Iterative OLS and PHD 1479
3.4 Iterative SIR and SAVE 1480
3.5 Iterative CUME and CUVE 1480
4 Simulated comparisons and examples 1481

4.1 Simulated comparisons 1481
4.2 Choosing K and the number of iterations 1486
4.3 Examples 1488
5 Conclusions 1491
Acknowledgements 1492
References 1492

1. Introduction

In regression analysis, one of the challenges faced with multi-dimensional data sets is graphical visualization to allow the relationship(s) between response and predictor variables to be detected. Within the past 20 years, in order to combat this challenge, several dimension reduction techniques that are very simple to implement have been introduced, including Sliced Inverse Regression (SIR) [16], Sliced Average Variance Estimation (SAVE) [8] and Principal Hessian Directions (PHD) [18]. Whilst classical slicing estimation methods such as SIR and SAVE require dividing the response into a discrete number of slices, more recent methods have been proposed which negate the need for slicing, such as Discretization-Expectation Estimation [27], Cumulative Mean Estimation (CUME), Cumulative Variance Estimation (CUVE) and Cumulative Directional Regression (CUDR) [26]. Consider a univariate response variable $Y \in \mathbb{R}$ and p-dimensional predictor vector $\mathbf{x} = [X_1, \ldots, X_p]^\top \in \mathbb{R}^p$. Then we can define the K-index model by

$$y = f\left(\boldsymbol{\beta}_1^\top \mathbf{x}, \ldots, \boldsymbol{\beta}_K^\top \mathbf{x}, \epsilon\right), \tag{1}$$

where $K < p$, f is an unknown *link function*, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are linearly independent p-dimensional column vectors and ϵ is an error term independent of $\mathbf{x}$. Define $\mathcal{S}$ to be the span of $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and note that replacing $\mathbf{x}$ with $\boldsymbol{\beta}_1^\top \mathbf{x}, \ldots, \boldsymbol{\beta}_K^\top \mathbf{x}$ reduces the dimension of the model, since the dimension of the latter is $K < p$. Li [16] coined $\mathcal{S}$ the effective dimension reduction (e.d.r.) space whose elements are e.d.r. directions. For identifiability purposes we will assume throughout that $\mathcal{S}$ is a central dimension reduction (CDR) subspace, which Cook [6] defined to be the intersection of all dimension reduction subspaces.

It is the aim of dimension reduction methods to find a basis for $\mathcal{S}$. However, existing methodologies are not without their limitations. To combine the noted strengths of various methods, several authors have promoted the combination of methods, in particular when one or more of the methods are restricted to only find a partial basis for $\mathcal{S}$ (see, for e.g., [17, 25, 28, 23]). Methods that find the full basis for $\mathcal{S}$ are said to have the "exhaustiveness" property, which was rigorously defined in [15].

The purpose of this article is to propose a new way of combining two (or potentially more) dimension reduction methods by using each one iteratively. This approach ensures that second (or further) iterations only return information regarding $\mathcal{S}$ that has not already been found. Section 2 discusses some existing methods and argues the advantages and disadvantages of each. Section

3 describes the theory and implementation of the iterative method of combining dimension reduction methods proposed in this article. The success of the iterative approach is assessed at the sample level in Section 4, where it is compared with various existing methods via simulations and examples. Concluding remarks are given in Section 5.

2. Dimension reduction

The use of dimension reduction methods to be implemented within this paper is not limited to the setting of a continuous Y. However, when Y is discrete it is difficult to conceptualize the role of the error term in (1). The model can be further generalized (see, for e.g., [6]) to state that Y and $\mathbf{x}$ are such that, for $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$,

$$Y \perp\!\!\!\perp \mathbf{x} | \mathbf{B}^\top \mathbf{x} \tag{2}$$

where $\perp\!\!\!\perp$ denotes independence. This is equivalent to requiring that Y depends on $\mathbf{x}$ only through $\boldsymbol{\beta}_1^\top \mathbf{x}, \ldots, \boldsymbol{\beta}_K^\top \mathbf{x}$ whilst avoiding the necessity of an error term.

Since the link function is unknown, we cannot uniquely determine the unique e.d.r. directions given by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. However, dimension reduction can still be achieved via $\boldsymbol{\gamma}_1^\top \mathbf{x}, \ldots, \boldsymbol{\gamma}_K^\top \mathbf{x}$ for any $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ such that the $\text{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K) = \text{Span}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$. In other words, any basis for $\mathcal{S}$ will suffice. Given a basis $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ for $\mathcal{S}$, a plot of Y versus $\boldsymbol{\gamma}_1^\top \mathbf{x}, \ldots, \boldsymbol{\gamma}_K^\top \mathbf{x}$ is called a Sufficient Summary Plot (SSP, see [6]) which can be used to explore the relationship between Y and $\mathbf{x}$ in the lower dimensional setting. At the sample level and for $\{y_i, \mathbf{x}_i\}_{i=1}^n$ denoting n sample realizations of $(Y, \mathbf{x})$ and $\widehat{\boldsymbol{\gamma}}_1, \ldots, \widehat{\boldsymbol{\gamma}}_K$ an estimated basis for $\mathcal{S}$, an Estimated SSP (ESSP) is constructed by plotting the y_i's versus the $\widehat{\boldsymbol{\gamma}}_1^\top \mathbf{x}_i, \ldots, \widehat{\boldsymbol{\gamma}}_K^\top \mathbf{x}_i$'s.

We now briefly discuss several dimension reduction techniques that will be the focus of our iterative approach. For simplicity we discuss them only at the population level (i.e. for a random $Y, \mathbf{x}$). However, all are moment based and therefore sample estimates follow simply.

2.1. Ordinary least squares

Whilst Ordinary Least Squares (OLS) was intended for use in the multiple linear regression framework, it is also useful in the dimension reduction setting. Let $\boldsymbol{\beta}_{ols} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{xy}$ denote the OLS slope vector where $\boldsymbol{\Sigma}_{xy} = \text{Cov}(\mathbf{x}, Y)$ is the covariance between $\mathbf{x}$ and Y. For $K = 1$ and an additive error in (1), Brillinger [2] showed that when $\mathbf{x}$ is multivariate normal, $\boldsymbol{\beta}_{ols} = c\boldsymbol{\beta}_1$ for a $c \in \mathbb{R}$. The conditions for which this result holds were generalized by Li & Duan [19]. They relaxed the normality condition for $\mathbf{x}$ and required only that $E(\mathbf{x} | \boldsymbol{\beta}_1^\top \mathbf{x})$ is linear in $\boldsymbol{\beta}_1^\top \mathbf{x}$. An additive error term was also no longer necessary.

In the case of a general K in (1), consider the following condition proposed by Li [16]:

Condition 2.1. $E\left(\mathbf{x}|\mathbf{B}^\top\mathbf{x}\right)$ *is linear in* $\mathbf{B}^\top\mathbf{x}$

which holds when $\mathbf{x}$ is elliptically symmetrically distributed. However, there are other situations which satisfy Condition 2.1 and Hall & Li [11] showed that it often approximately holds for high dimensional $\mathbf{x}$.

Condition 2.1 is equivalent to $E(\mathbf{x}|\mathbf{B}^\top\mathbf{x}) = \boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^\top\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^\top(\mathbf{x}-\boldsymbol{\mu})$ (see, for e.g., the proof to Lemma 1 in [21]). Let $\mathbf{u}$ be orthogonal to $\mathcal{S}$. Under Condition 2.1,

$$\begin{aligned}\mathbf{u}^\top\boldsymbol{\Sigma}^{-1}\mathrm{Cov}(\mathbf{x},Y) =& \mathbf{u}^\top\boldsymbol{\Sigma}^{-1}E\left\{E\left[(\mathbf{x}-\boldsymbol{\mu})(Y-E(Y))|\mathbf{B}^\top\mathbf{x}\right]\right\}\\ =& \mathbf{u}^\top\boldsymbol{\Sigma}^{-1}E\left\{(Y-E(Y))E\left[(\mathbf{x}-\boldsymbol{\mu})|\mathbf{B}^\top\mathbf{x}\right]\right\} = 0.\end{aligned}$$

Hence, under Condition 2.1 and the model in (1), $\mathbf{u}^\top\boldsymbol{\beta}_{ols} = 0$ for all $\mathbf{u}$ orthogonal to $\mathcal{S}$ so that $\boldsymbol{\beta}_{ols} \in \mathcal{S}$. As a result, even when $K > 1$, OLS can still be used to extract part of $\mathcal{S}$. Further discussion and results on this in a more general setting (that includes OLS) can be found on pages 143-147 of Cook [6]. This is an important point when we discuss opportunities for iterative use of dimension reduction methods later.

2.2. Sliced inverse regression

Unlike OLS, Sliced Inverse Regression [16] (SIR) can, although is not guaranteed to, return an entire basis for $\mathcal{S}$ when $K > 1$. Under Condition 2.1, Li [16] showed that eigenvectors corresponding to non-zero eigenvalues of $\mathrm{Cov}[E(\mathbf{x}|Y)]$ are elements of $\boldsymbol{\Sigma}\mathcal{S}$ (which we define as the basis for the span of the $\boldsymbol{\Sigma}\boldsymbol{\beta}_i$'s). Let $S_1, S_2, \ldots, S_H$ denote non-overlapping subranges of Y such that $\bigcup_{h=1}^{H} S_h = \mathrm{range}(Y)$. Li circumvented the challenge faced with determining $E(\mathbf{x}|Y)$ by utilizing 'slicing' which is equivalent to discretizing Y according to which S_h it belongs to. This slicing strategy means that $E(\mathbf{x}|Y)$ is approximated by the slice means given as $E(\mathbf{x}|Y \in S_h) = \boldsymbol{\mu}_h$, $(h = 1, \ldots, H)$. The choice of p_h's is up to the researcher although it is common (and straightforward) to choose equally probable slices such that each $p_h = 1/H$. In practice this is akin to choosing an (approximately) equal number of observations per slice.

Let $\boldsymbol{\Gamma} = \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ denote the basis for $\mathcal{S}$ returned by SIR. When SIR operates on the standardized predictor $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$ and is followed by a re-standardization back to the original scale, then $\boldsymbol{\Gamma}^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \mathbf{I}_K$ such that the dimension reduced predictors $\boldsymbol{\gamma}_1^\top\mathbf{x}, \ldots, \boldsymbol{\gamma}_K^\top\mathbf{x}$ are uncorrelated and each have unit variance. The SIR algorithm is therefore

Step 1 For $\mathbf{p} = [p_1, \ldots, p_H]$ and $\boldsymbol{\mu}_{z,h} = E(\mathbf{Z}|Y \in S_h)$, determine

$$\mathbf{V}_\mathbf{p} = \sum_{h=1}^{H} p_h\boldsymbol{\mu}_{z,h}\boldsymbol{\mu}_{z,h}^\top.$$

Step 2 Return $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_1, \ldots, \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_K$ as a basis for $\mathcal{S}$ where $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ are the eigenvectors corresponding to the K largest eigenvalues of $\mathbf{V}_\mathbf{p}$.

Under Condition 2.1 and the model in (2), $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta}_i \in \mathcal{S}$ if λ_i is non-zero [16, 6]. Hence, provided $\text{rank}(\mathbf{V_p}) = K$, SIR returns a basis for $\mathcal{S}$ under these conditions. However, there are two notable limitations which can cause $\text{rank}(\mathbf{V_p}) < K$, both of which are are important to this paper. Firstly, as illustrated by Cook & Weisberg [8], for model types which exhibit some symmetric dependency structure about the mean of $\mathbf{x}$, SIR will fail to find an entire basis for $\mathcal{S}$ (see, for e.g.,[16, 8]). Secondly, since $p_1\boldsymbol{\mu}_{z,1} + p_2\boldsymbol{\mu}_{z,2} + \ldots + p_H\boldsymbol{\mu}_{z,H} = \mathbf{0}$, the $\boldsymbol{\mu}_{z,h}$'s are linearly dependent so that the maximum rank of $\mathbf{V_p}$ is $H-1$. Whilst for a continuous Y one can choose H large enough so that $H-1 \geq K$, when Y is discrete the response discretization is not necessary and the number of slices H is equal to the number of unique elements in the response space. For example, if Y is binary then there are $H = 2$ slices and the maximum rank of $\mathbf{V_p}$ is one. Both of these limitations mean that for some model types, SIR may only find part of a basis for $\mathcal{S}$.

2.3. Sliced average variance estimation

Consider the following condition:

Condition 2.2. *Cov* $\left(\mathbf{x}|\mathbf{B}^\top\mathbf{x}\right)$ *is constant.*

Recalling the slicing approach used by SIR and discussed in Section 2.2, Cook & Weisberg [8] showed that if Conditions 2.1 and 2.2 hold (which are both satisfied when $\mathbf{x}$ is normal), then slice covariances also contain information regarding $\mathcal{S}$ (Li and Wang [14] also noted that SAVE holds the exhaustiveness property if $\mathbf{x}|Y$ is normally distributed). Cook & Weisberg therefore introduced SAVE, which may be implemented in a similar manner to SIR but where Step 1 in the SIR algorithm becomes, for SAVE,

Step 1 For $\mathbf{p} = [p_1, \ldots, p_H]$ and $\boldsymbol{\Sigma}_{z,h} = \text{Cov}(\mathbf{Z}|Y \in S_h)$, determine

$$\mathbf{M_p} = \sum_{h=1}^{H} p_h(\mathbf{I}_p - \boldsymbol{\Sigma}_{z,h})^2$$

and the eigen-analysis in Step 2 is conducted on $\mathbf{M_p}$ to obtain a basis for $\mathcal{S}$ after re-standardization.

SAVE does not suffer the same restrictions with respect to symmetric dependency or rank deficiencies for a discrete response that SIR does. Therefore, provided the additional Condition 2.2 also holds, SAVE can be a useful approach for returning a basis for $\mathcal{S}$.

2.4. Principal Hessian directions

For $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, Li [18] used Stein's Lemma [24] and properties of the Hessian matrix to introduce PHD which may also be used to derive a basis for $\mathcal{S}$. As with SAVE, the algorithm for PHD is similar to SIR where

Step 1 For $\mu_Y = E(Y)$, determine the weight-covariance matrix

$$\boldsymbol{\Sigma}_{yzz} = E\left[(Y - \mu_Y)\,\mathbf{Z}\mathbf{Z}^\top\right] \tag{3}$$

and, again, the eigen-analysis in Step 2 is now conducted on $\boldsymbol{\Sigma}_{yzz}$ to obtain a basis for $\mathcal{S}$ after re-standardization.

One advantage of PHD over SIR and SAVE is that a slicing parameter need not be chosen. Additionally, whilst the normality condition for $\mathbf{x}$ seems restrictive, PHD is still applicable if Conditions 2.1 and 2.2 hold, although, as noted by Cook [5], the connection with the Hessian matrix may be lost. As stated by Li [18], one restriction of PHD is that due to the nature of the Hessian matrix, PHD is not expected to return linear components.

Li [18] also highlighted that when linear functions of $\mathbf{x}$ are added or subtracted from Y then, under the normality condition for $\mathbf{x}$, the Hessian matrix does not change and at the population level PHD is not effected. This lead to another version of PHD that utilizes

$$\boldsymbol{\Sigma}_{rzz} = E\left[R\mathbf{Z}\mathbf{Z}^\top\right] \tag{4}$$

where $R = Y - \mu_Y - \boldsymbol{\beta}_{ols}^\top(\mathbf{x} - \boldsymbol{\mu})$ is the OLS residual. This version of PHD is typically preferred (see, for e.g., [5]) and, for many model types, estimator variability is smaller for this residuals-based approach when compared to the standard PHD approach [22]. We note that this residuals approach simply involves a transformation of Y using the OLS residual before the implementation of PHD. This is distinguished from the method we later propose in Section 3.3, which instead involves first implementing OLS in the dimension reduction sense to obtain the first component of $\mathcal{S}$ (rather than to simply transform Y), followed by the implementation of residuals-based PHD.

2.5. Combined approaches

In the rejoinder to his original paper which introduced SIR, Li [17] introduced SIRII which, like SAVE, utilized slice covariances. In doing so Li introduced SIRII_α, for which a convenient notation is $\text{SIRII}_\alpha = (1-\alpha)\text{SIR}^2 + \alpha(\text{SIRII})$, which should be read as sum of $(1-\alpha)$ times the SIR matrix squared (i.e. $\mathbf{V}_\mathbf{p}^2$) and α times the SIRII matrix which is based on sliced covariances. In a similar fashion, Ye & Weiss [25] introduced the combination of $(1-\alpha)$SIR+αPHD as well as $(1-\alpha)$SIR+αSAVE. Zhu et al. [28] analysed these methods further and also introduced variations on the combination of $(1-\alpha)$SIR+αSAVE. As a result of their analysis, the hybrid method they recommended was $(1-\alpha)$SIR+αSAVE with a parameter value of $\alpha = 0.5$.

2.6. Cumulative slicing procedures

Whilst the slicing approaches such as SIR and SAVE are simple to utilize, a possible deficiency in the continuous response case is that they may be sensitive

to the number of slices chosen. Zhu et al. [26] extended these methods by using the idea of cumulative slicing. Whilst SIR extracts information regarding $\mathcal{S}$ using slice means given as $E(\mathbf{x}|Y \in S_h)$ and SAVE via the slice covariances denoted $\mathrm{Cov}(\mathbf{x}|Y \in S_h)$, $(h = 1, \ldots, H)$, Cumulative Mean Estimation (CUME) and Cumulative Variance Estimation (CUVE) instead use $E[\mathbf{x}I(Y \leq \widetilde{y})]$ and $\mathrm{Cov}[\mathbf{x}I(Y \leq \widetilde{y})]$ for all $\widetilde{y} \in \mathbb{R}$ and where $I(\cdot)$ is the indicator function. Like SIR, CUME requires that Condition 2.1 holds and is similarly limited with respect to detecting all components when symmetric dependency structure is evident. Similarly to SAVE, CUVE requires that both Conditions 2.1 and 2.2 hold.

In the same way as for SIR, SAVE and PHD, we consider CUME and CUVE initially on the standardized scale with a re-standardization back to the original scale using $\boldsymbol{\Sigma}^{-1/2}$ following the eigen-decomposition step. For CUME, we then have:

Step 1: Determine $E\left\{E[\mathbf{Z}I(Y \leq \widetilde{y})]E[\mathbf{Z}I(Y \leq \widetilde{y})]^{\top}\right\}$

and for CUVE:

Step 1: For $\widetilde{p} = P(Y \leq \widetilde{y})$, determine $E\left\{[\widetilde{p}\mathbf{I}_p - \mathrm{Cov}(\mathbf{Z}I(Y \leq \widetilde{y}))]^2\right\}$.

The approaches also allow for the possibility of weighting with respect to Y although we are specifically dealing with the case of equal weighting as recommended by Zhu et al. [26], which are shown above. In addition to CUME and CUVE, Zhu et al. also provide a cumulative slicing extension to the directional regression proposed by Li and Wang [14]. Complementally to the CUME and CUVE approaches, Cumulative Directional Regression (CUDR) uses $E[(\widetilde{\mathbf{x}} - \mathbf{x})(\widetilde{\mathbf{x}} - \mathbf{x})^{\top} I(Y \leq y)I(\widetilde{Y} \leq \widetilde{y})]$ which contains information regarding $\boldsymbol{\Sigma}\mathcal{S}$ where $(\widetilde{Y}, \widetilde{x})$ is an independent copy of $(Y, \mathbf{x})$. CUME, CUVE and CUDR are all moment-based approaches and so can be employed conveniently in practice.

3. Iterative use of inverse regression methods

Before we discuss a simple iterative approach for dimension reduction, we will consider a motivating example that highlights how SIR can be a very good estimator of a partial basis for $\mathcal{S}$.

3.1. A motivating example

Consider the $K = 2$ model

$$Y = \left(\boldsymbol{\beta}_1^{\top}\mathbf{x}\right)^2 + \left(\boldsymbol{\beta}_2^{\top}\mathbf{x}\right)^3 \tag{5}$$

where $\mathbf{x} \sim N_{10}(\mathbf{0}, \mathbf{I}_{10})$, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^{\top}$ and $\boldsymbol{\beta}_2 = (0, 1, 0, \ldots, 0)^{\top}$. As discussed by Cook & Weisberg [8] and Li [16], we would expect SIR to fail at finding the $\left(\boldsymbol{\beta}_1^{\top}\mathbf{x}\right)^2$ part of the model due to symmetric dependency around the mean of $\mathbf{x}$. However, we do not expect SAVE to have trouble finding this component.

For this example we will consider simulating $n = 120$ observations according to this model. Let $\mathbf{X}_n$ denote the 120×10 matrix whose ith row is $\mathbf{x}_i$—the simulated regressor vector for the ith observation. Let $\widehat{\boldsymbol{\Gamma}}_{sir}$ and $\widehat{\boldsymbol{\Gamma}}_{save}$ denote

TABLE 1
Average canonical correlations for 1000 trials of SIR and SAVE for $n = 120$ observations generated from the model in (5). Standard deviations are in parentheses

	SIR	SAVE
1st canonical correlation	0.978 (0.014)	0.939 (0.042)
2nd canonical correlation	0.258 (0.195)	0.458 (0.285)

the $K = 2$ basis estimates for $\mathcal{S}$ that are returned by SIR and SAVE respectively. Since the dimension reduced regressors are of primary importance and it is a basis for $\mathcal{S}$ that is to be targeted, we will report the canonical correlations between each of the estimated dimension reduced regressors and $\mathbf{XB}$ (the dimension reduced regressors with $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$). We have also chosen $H = 5$ equally probable slices.

We provide the average canonical correlations in Table 1 where standard deviations of these correlations are in parentheses. We can see that SIR is highly capable of finding one component with an average first canonical correlation of 0.978 and a small standard deviation of only 0.014. However, it typically performs poorly with respect to the second element of the basis where the average canonical correlation is only 0.258. SAVE also performs very well with the first component (average first canonical correlation of 0.939) but not as well as SIR. The standard deviation for the SAVE correlation is also three times that of SIR indicating higher variability in estimation. As expected, SAVE performs better than SIR when estimating the second component.

In conclusion for this motivating example, we see that SIR is capable of estimating a partial basis extremely well for this model. It would therefore be beneficial to use this partial estimate with another use of SAVE.

3.2. Theory

We now discuss the iterative usage of dimension reduction methods that preserves a good partial estimate of a basis for $\mathcal{S}$ found by the first method. Given the methods discussed in Section 2, a simple approach to achieve this is to remove the already estimated component from the matrix derived for the second dimension reduction method in Step 1.

Proposition 3.1. *Let $\mathbf{C}$ be a $p \times p$ symmetric matrix whose eigenvectors corresponding to non-zero eigenvalues are elements of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$. Furthermore, let $\mathbf{P}$ be a projection matrix onto $\boldsymbol{\Sigma}^{1/2}\mathcal{S}_1 \subset \boldsymbol{\Sigma}^{1/2}\mathcal{S}$. Let $\mathcal{S}_1'$ be the complement of $\mathcal{S}_1$. Then the eigenvectors corresponding to the non-zero eigenvalues of $(\mathbf{I}_p - \mathbf{P})\,\mathbf{C}\,(\mathbf{I}_p - \mathbf{P})$ are elements of $\boldsymbol{\Sigma}^{1/2}\mathcal{S} \cap \boldsymbol{\Sigma}^{1/2}\mathcal{S}_1'$ (the basis whose elements are all in $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ but not in $\boldsymbol{\Sigma}^{1/2}\mathcal{S}_1$).*

Proof. Since $\mathbf{P}$ is a projection matrix onto a subspace $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$, then $\mathbf{I}_p - \mathbf{P}$ is a projection matrix onto the complement of this space. Hence, any eigenvectors corresponding to nonzero eigenvalues must still be elements of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ although they must also be orthogonal to $\boldsymbol{\Sigma}^{1/2}\mathcal{S}_1$. □

Proposition 3.1 is useful because it provides a simple means for which to remove from estimation the already determined components of the dimension reduction subspace. The result is general, and there exist many possible combinations of existing dimension reduction methods that can be used iteratively. We will focus our attention on three particular iterative combinations where we have chosen the methods due mainly to (i) the first method having proven to be a useful estimator in many applications though also having noted limitations and (ii) the second method being a natural pairing with the first and one which seeks to overcome the aforementioned restrictions.

We choose to operate firstly on the **z**-scale (i.e. focus on $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$) and then to re-standardize to the **x**-scale to preserve some convenient characteristics of the dimension reduction methods discussed earlier. For example, if $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K$ are a basis for $\mathcal{S}$ returned by SIR, SAVE or PHD, then $\text{Var}(\boldsymbol{\gamma}_k^\top \mathbf{x}) = 1$ and $\text{Cov}(\boldsymbol{\gamma}_k^\top \mathbf{x}, \boldsymbol{\gamma}_j^\top \mathbf{x}) = 0$ $(k \neq j)$. That is, the dimension reduced regressors are scaled to have variance 1 and they are also mutually uncorrelated. By removing the component on the **z**-scale for the iterative approach and then re-standardizing with respect to $\boldsymbol{\Sigma}^{-1/2}$ (as is done for SIR, SAVE and PHD) we retain these features for the final dimension reduction subspace basis.

3.3. Iterative OLS and PHD

As in Section 2.1, let $\boldsymbol{\beta}_{ols}$ denote the least squares slope vector which, under the model in (1) and Condition 2.1, is an element of $\mathcal{S}$. For simplicity, let $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L$ denote the PHD directions where under the assumption of a normal **x** (or the less restrictive combination of Conditions 2.1 and 2.2) and the model in (1), $\text{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L) \subseteq \mathcal{S}$. Given that OLS can only, at most, find one direction in the dimension reduction subspace, our intention here is to use PHD conditionally on the implementation of OLS. To remove the OLS component from the basis to be returned by PHD, we define the associated projection matrix for OLS as

$$\mathbf{P}_{ols} = \left(\boldsymbol{\beta}_{ols}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_{ols}\right)^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}_{ols} \boldsymbol{\beta}_{ols}^\top \boldsymbol{\Sigma}^{1/2} \tag{6}$$

which projects onto the subspace of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ spanned by $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_{ols}$.

Let the PHD matrix whose eigenvectors corresponding to non-zero eigenvalues are elements of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ be denoted $\boldsymbol{\Sigma}_{\cdot zz}$ which is equal to either $\boldsymbol{\Sigma}_{yzz}$ or $\boldsymbol{\Sigma}_{rzz}$ as denoted in (3) and (4) respectively. Therefore, to remove the component already retrieved by OLS, Step 1 of the PHD algorithm becomes

Step 1 Determine $\left(\mathbf{I}_p - \mathbf{P}_{ols}\right) \boldsymbol{\Sigma}_{\cdot zz} \left(\mathbf{I}_p - \mathbf{P}_{ols}\right)$.

Step 2 then involves an eigen-analysis of this matrix and re-standardized eigenvectors corresponding to nonzero eigenvalues to form a basis for $\mathcal{S}$ when combined with the original OLS direction. Let $\boldsymbol{\Gamma}$ be this basis. Then choosing a re-scaled OLS direction of $\left(\boldsymbol{\beta}_{ols}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_{ols}\right)^{-1/2} \boldsymbol{\beta}_{ols}$ results in $\boldsymbol{\Gamma}^\top \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \mathbf{I}_K$ so that each dimension reduced predictor has unit variance and the dimension reduction predictors are mutually uncorrelated (note, however, that simply choosing $\boldsymbol{\beta}_{ols}$

still results in uncorrelated dimension reduced regressors). We will refer to this approach as PHD|OLS. We note that PHD|OLS is quite different from the Iterative Hessian Directions (IHT) method proposed by Cook and Li [7]. Although both methods are hybrids of OLS and PHD, PHD|OLS uses OLS and PHD separately by retrieving a component by OLS first and then retrieving further components using PHD conditional on the OLS component already obtained. By contrast, IHT retrieves a basis for $\mathcal{S}$ based on an eigen decomposition of a matrix constructed using the OLS direction and powers of the Hessian matrix. Lue et al. [20] also use OLS in conjunction with r-based PHD, but in a censored survival regression setting.

3.4. Iterative SIR and SAVE

SIR and SAVE make a natural pairing with SIR utilizing slice means and SAVE slice covariances. When Y is continuous, the same slicing strategy can be implemented for both although this isn't strictly necessary. For the iterative approach we choose to use SIR first since (i) simulations have shown that it has very good estimation properties for a wide choice of models and (ii) there are some known and discussed limitations involving some model types where only part of a basis for $\mathcal{S}$ is achievable. Let $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_L$ denote a partial basis for $\mathcal{S}$ returned by SIR. Then the projection matrix onto the associated subspace of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ is equal to

$$\mathbf{P}_{sir} = \boldsymbol{\Sigma}^{1/2} \sum_{l=1}^{L} \boldsymbol{\gamma}_l \boldsymbol{\gamma}_l^{\top} \boldsymbol{\Sigma}^{1/2}. \tag{7}$$

For this approach, which we will refer to as SAVE|SIR, Step 1 becomes

Step 1 Determine $(\mathbf{I}_p - \mathbf{P}_{sir})\, \mathbf{M}_{\mathbf{p}}\, (\mathbf{I}_p - \mathbf{P}_{sir})$

and Step 2 is applied to this matrix. For simplicity, let $\boldsymbol{\Gamma}$ denote the basis for $\mathcal{S}$ which consists of the original SIR directions and the additional SAVE directions from the iterative step, which are the elements corresponding to re-standardized eigenvectors corresponding to non-zero eigenvalues of the matrix in Step 2 above. Then, again, $\boldsymbol{\Gamma}^{\top}\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \mathbf{I}_K$ resulting in unit variance dimension reduced regressors that are mutually uncorrelated.

3.5. Iterative CUME and CUVE

The iterative approach for CUME and CUVE is almost identical to that taken for SIR and SAVE. Since CUME requires fewer conditions on $\mathbf{x}$ than what is required for CUVE, but where CUME may be restricted due to symmetric dependency, we focus on the application of CUME to obtain a partial basis followed by CUVE to find the remaining elements of a basis for $\mathcal{S}$. Here we define $\mathbf{P}_{cume}$ as the projection matrix for the partial basis of $\boldsymbol{\Sigma}^{1/2}\mathcal{S}$ obtained by CUME. Application of CUVE then follows after a pre- and post-multiplication of $(\mathbf{I}_p - \mathbf{P}_{cume})$ with respect to the matrix shown in Step 1 for CUVE.

4. Simulated comparisons and examples

4.1. Simulated comparisons

In this Section, we compare the performance of SIR, SAVE, $(1-\alpha)$ SIR+αSAVE (which, for simplicity, we will refer to as SIR+SAVE), SAVE|SIR, PHD|OLS, CUDR and CUVE|CUME using simulated data. Note that SIR+SAVE is equivalent to SIR when $\alpha = 0$ and SAVE when $\alpha = 1$. All models considered are $K = 2$ models. Simulations were run 500 times for each model and we used the canonical correlations between $\widehat{\boldsymbol{\gamma}}_1^\top \mathbf{x}, \ldots, \widehat{\boldsymbol{\gamma}}_K^\top \mathbf{x}$ and $\boldsymbol{\beta}_1^\top \mathbf{x}, \ldots, \boldsymbol{\beta}_K^\top \mathbf{x}$ as the method of assessment. Similar methods of assessment are commonly used in the context of dimension reduction analysis (see, for example, [16] and [28]).

The first two models we consider have a continuous response. For these models we look at the first two canonical correlations, denoted r_1 and r_2, separately, as well as looking at averages. Since we will be looking at $K = 2$ models where in many cases some methods will be expected to estimate one direction very well and the other poorly, seeing r_1 and r_2 separately will provide useful insight.

For $\mathbf{x} \sim N_{10}(\mathbf{0}, \mathbf{I}_{10})$ and $\epsilon \sim N(0,1)$, the first two models are

Model 1 For $\boldsymbol{\beta}_1 = [1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [0, 1, 0, \ldots, 0]^\top$,

$$Y = \left(\boldsymbol{\beta}_1^\top \mathbf{x}\right)^3 + \left(\boldsymbol{\beta}_2^\top \mathbf{x}\right)^2 + \epsilon$$

Model 2 For $\boldsymbol{\beta}_1 = [1, 2, -3, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [1, 1, 0, -2, 0, \ldots, 0]^\top$,

$$Y = \sin\left(0.5\boldsymbol{\beta}_1^\top \mathbf{x}\right) + \cos\left(0.5\boldsymbol{\beta}_2^\top \mathbf{x}\right) + 0.3\epsilon \tag{8}$$

Figure 1 shows the results of the simulation for Model 1 for 500 trials. The first two figures show the mean canonical correlations (first and second) whilst the third is the mean of the average of the first two canonical correlations. The table displays the standard deviations for the average canonical correlations. Six methods are compared; namely SAVE, SAVE|SIR, PHD|OLS, SIR+SAVE, CUDR and CUVE|CUME. Both the residuals-based and standard PHD approaches were considered, however only the residuals-based PHD results are shown, as they were typically superior for this example. For SIR+SAVE, three values of α were considered (0.2, 0.5 and 0.8), however for the sake of brevity only the SIR+SAVE results with $\alpha = 0.2$ are reported, as its performance was superior to SIR+SAVE with α =0.5 and 0.8. We have chosen $n = 50, 100, 200$ and 400 and $H = 2, 5$ and 10 equally probable choices for SIR and SAVE.

Model 1 was used for simulations in [28] because it has both a symmetric $(\boldsymbol{\beta}_2^\top \mathbf{x})^2$ and asymmetric $(\boldsymbol{\beta}_1^\top \mathbf{x})^3$ element, so we would expect SIR to do well at identifying the asymmetric element and SAVE to do well at finding the symmetric element. Figure 1 shows that PHD|OLS and CUVE|CUME's results are better than or equal to those of any of the other methods for all choices of n for both r_1 and r_2. SAVE|SIR is the next best performer, followed by SIR+SAVE and CUDR. The canonical correlations for SAVE are consistently the lowest for all choices of n and H and for both directions. SAVE also shows relative

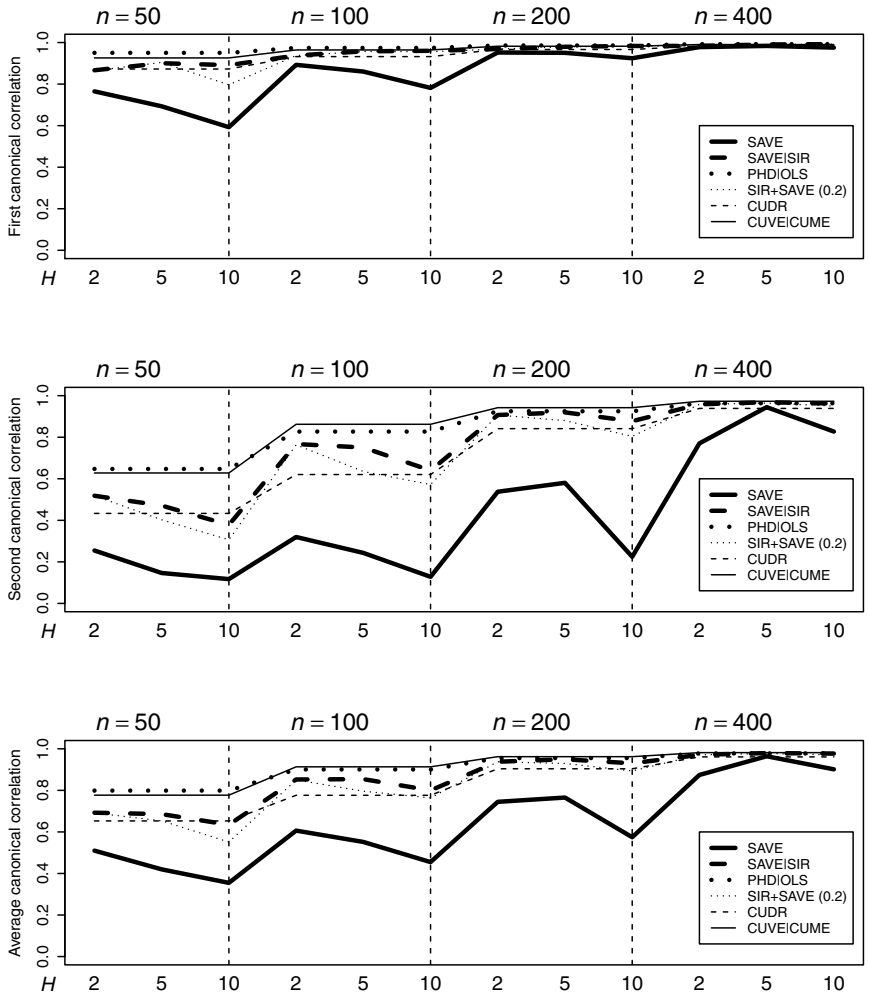

Table of Standard Deviations

n	SAVE	SAVE|SIR	PHD|OLS	S+S $\alpha = 0.2$	CUDR	CUVE|CUME
50	0.12	0.14	0.14	0.13	0.14	0.14
100	0.1	0.1	0.1	0.13	0.13	0.08
200	0.14	0.03	0.05	0.06	0.08	0.02
400	0.03	0.01	0.01	0.01	0.02	0.01

FIGURE 1. *These figures show the mean first, second and average canonical correlations for 500 iterations of data simulated from Model 1 for four choices of n (50, 100, 200 and 400) and three values of H (2, 5 and 10) for equally probable slices for SIR and SAVE. S+S refers to SIR+SAVE. Standard deviations for the average canonical correlations are shown in the table.*

sensitivity to the choice of H when compared with SAVE|SIR and SIR+SAVE. PHD|OLS, CUDR and CUVE|CUME are not affected by the choice of H, since slicing is not required for these methods. A summary of standard deviations are shown within the figure for the average canonical correlations when $H = 5$. The standard deviations show that, in general, the variability of all methods decreases as n increases.

Similarly to the analysis for Model 1, Figure 2 shows the results of the simulation for Model 2. This model was chosen because, as for Model 1, the model contains a symmetric and an asymmetric element. Again, CUVE|CUME and PHD|OLS are the best performers for this model, and they produce superior results for all choices of n and H and for both directions. Next are SAVE|SIR and SIR+SAVE which provide similar results, followed by CUDR and SAVE respectively. Models 1 and 2 show that for all chosen values of n, H and α, the results of CUVE|CUME, SAVE|SIR and PHD|OLS are better than or equal to those of SIR+SAVE and SAVE.

In the examples we have considered thus far, PHD|OLS and CUVE|CUME have performed exceptionally well. However, PHD|OLS can be sensitive to extremely large response values (in comparison to other responses), whereas slicing methods such as SIR, SAVE, CUME etc. can be robust to particularly large (or small) response values since the response plays a 'positioning role' only for the ordering of the regressors and we briefly highlight this here. Additionally, and in fairness to PHD|OLS, we also provide an example where PHD|OLS is the preferred approach in a setting where large responses are not a concern. The third model we therefore consider is from Zhu et al. [26] and is defined to be, for $\mathbf{x} \sim N_{15}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the covariance matrix with ijth element defined to be $\sigma_{ij} = 0.2^{|i-j|}$, and $\boldsymbol{\beta}_1 = [1, 1, 1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [1, 0, 0, 0, 1, 3, 0, \ldots, 0]^\top$,

Model 3

$$Y = \exp\left[0.2\left(\boldsymbol{\beta}_1^\top \mathbf{x} + 1\right)^3 + 0.2\left(1 + \left(\boldsymbol{\beta}_2^\top \mathbf{x}/2\right)^2\right) + 0.2\epsilon\right]. \tag{9}$$

We also reconsider Model 2 but with directions $\boldsymbol{\beta}_1 = [1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [0, 1, 0, \ldots, 0]^\top$. For brevity we report only the average canonical correlations and standard deviations for 500 trials, $p = 15$ and $n = 400$.

Table 2 shows that for Model 3, which contains extremely large response values in comparison to other responses, PHD|OLS performs poorly, achieving an average canonical correlation of just 0.465, compared to 0.949 and 0.968 for SAVE|SIR and CUVE|CUME respectively. On the other hand, PHD|OLS clearly outperforms the other two methods for modified Model 2, achieving an average canonical correlation of 0.968 compared to 0.82 for SAVE|SIR and 0.898 for CUVE|CUME.

We now consider a model which has a discrete binary response, which, for $\mathbf{x} \sim N_{10}(\mathbf{0}, \mathbf{I}_{10})$ and $\epsilon \sim N(0, 1)$, is defined to be

Model 4 For $\boldsymbol{\beta}_1 = [1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [0, 1, 0, \ldots, 0]^\top$,

$$Y = \begin{cases} 1, & \text{if } \boldsymbol{\beta}_1^\top \mathbf{x} < \left(\boldsymbol{\beta}_2^\top \mathbf{x}\right)^2 - 0.65 + 0.25\epsilon; \\ 2, & \text{otherwise.} \end{cases}$$

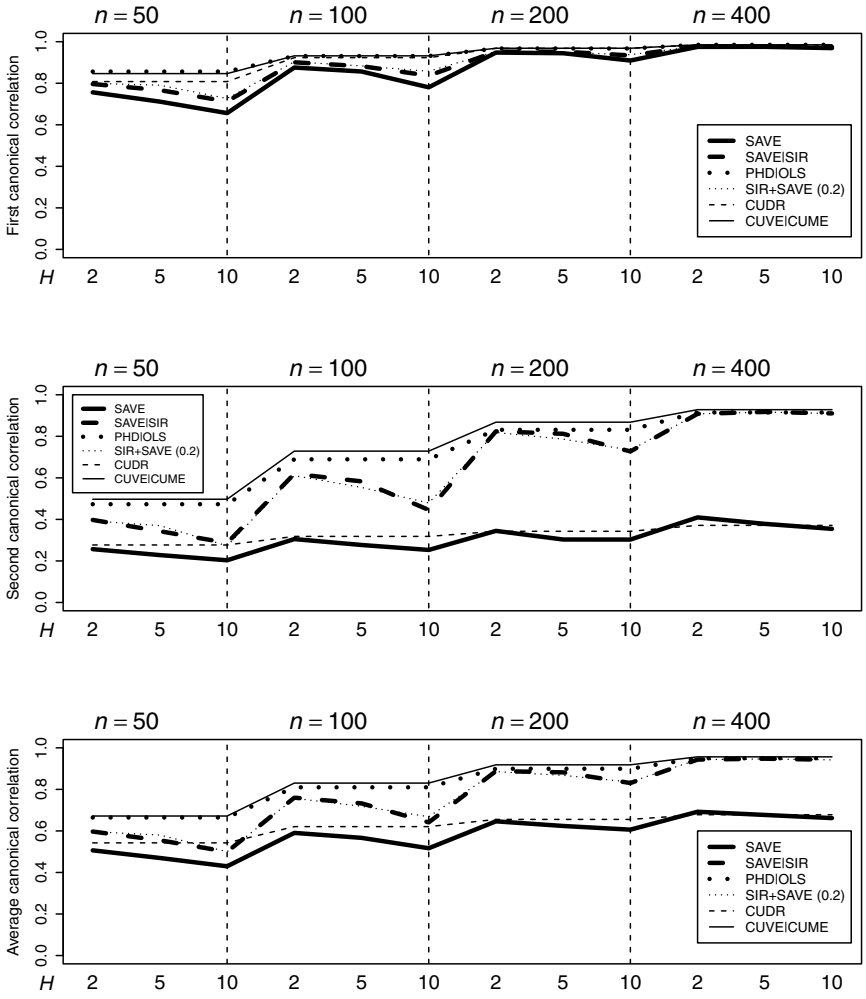

Table of Standard Deviations

n	SAVE	SAVE|SIR	PHD|OLS	S+S $\alpha = 0.2$	CUDR	CUVE|CUME
50	0.14	0.16	0.15	0.15	0.13	0.15
100	0.11	0.14	0.11	0.14	0.12	0.09
200	0.11	0.07	0.06	0.09	0.12	0.05
400	0.13	0.03	0.03	0.03	0.12	0.02

FIGURE 2. *These figures show the mean first, second and average canonical correlations for 500 iterations of data simulated from Model 2 for four choices of n (50, 100, 200 and 400) and three values of H (2, 5 and 10) for equally probable slices for SIR and SAVE. S+S refers to SIR+SAVE Standard deviations for the average canonical correlations are shown in the table.*

TABLE 2
*Average canonical correlations for 500 trials of PHD|OLS, SAVE|SIR and CUVE|CUME for $n = 400$ observations generated from Models 4 and a modified Model 2. *For Model 2, $\boldsymbol{\beta}_1 = [1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [0, 1, 0, \ldots, 0]^\top$. Standard deviations are in parentheses*

	PHD|OLS	SAVE|SIR	CUVE|CUME
Model 3	0.465 (0.084)	0.949 (0.036)	0.968 (0.017)
Model 2*	0.968 (0.015)	0.82 (0.124)	0.898 (0.085)

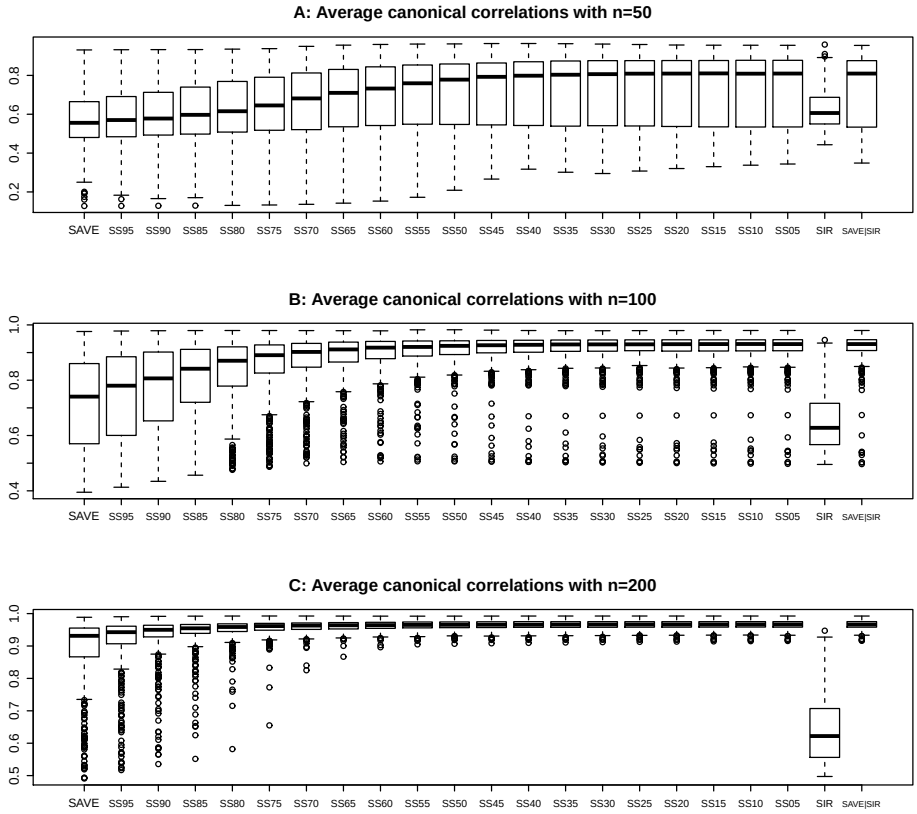

FIGURE 3. *Boxplots of the average canonical correlations for 500 iterations of data simulated from Model 4 for three choices of n (50, 100 and 200). The methods compared are SIR, SAVE, SAVE|SIR and SIR+SAVE with values of $\alpha = 0.95$ down to $\alpha = 0.05$ in increments of 0.05. SS95 refers to SIR+SAVE with $\alpha = 0.95$, SS90 is SIR+SAVE with $\alpha = 0.9$, etc.*

Since Model 4 is a binary response model, slicing occurs naturally. As such, we focus our analysis on the SIR and SAVE based approaches which are then naturally suited to this model. In Figure 3 we have chosen to show boxplots of the average canonical correlations so that we could highlight a particular commonality between SIR+SAVE and SAVE|SIR. The results for SIR and SAVE|SIR are on the right. SIR typically performs poorly due to its inability to target an entire basis for $\mathcal{S}$. SAVE does not perform well for small n, although its perfor-

mance improves with increasing n. The results of SAVE|SIR and SIR+SAVE are superior to those of SIR and SAVE, however the advantage of SAVE|SIR is that it does not require the choice of an additional parameter α. What is obvious, however, is that SIR+SAVE clearly improves as α decreases but also the results seem to also approach those of SAVE|SIR. We discuss this in the following remark.

Remark 4.1. As noted earlier in this section, SIR+SAVE is equivalent to SAVE when $\alpha = 1$ and equivalent to SIR when $\alpha = 0$. As such, we would expect that as α increases, the results of SIR+SAVE would approach those of SAVE, and that as α decreases, the results of SIR+SAVE would approach those of SIR. Figure 3 seems to support the former claim, however the opposite of the latter claim has occurred where as α decreases, the results seem to approach those of SAVE|SIR rather than SIR. This is limited to the binary response case and can be explained. In the binary case, the SIR matrix has exact rank one. Therefore, when the SAVE matrix is added, all of the additional information (other than that already found by SIR) is in that matrix even if α is small (but nonzero). A small α increases the chance of the SIR information being undisturbed by SAVE for the SIR+SAVE approach. That is, for small α the SIR matrix is prominent and the SIR information is likely to be utilized. For large α, the SAVE matrix becomes dominant and, as such, the SAVE information is likely to contribute the most to the estimated basis.

In Remark 4.1 we highlighted why SIR+SAVE seems to approach SAVE|SIR when Y is binary despite it being typically expected to approach SIR with decreasing α. To highlight this further we reconsider Model 1 (a continuous model) and also now consider a fifth model, which has a discrete ternary response, defined as

Model 5 For $\boldsymbol{\beta}_1 = [1, 0, \ldots, 0]^\top$ and $\boldsymbol{\beta}_2 = [0, 1, 0, \ldots, 0]^\top$,

$$Y = \begin{cases} 1, & \text{if } \boldsymbol{\beta}_1^\top \mathbf{x} < \boldsymbol{\beta}_2^\top \mathbf{x}^2 - 0.95 + 0.25\epsilon; \\ 2, & \text{if } \boldsymbol{\beta}_2^\top \mathbf{x}^2 - 0.95 + 0.25\epsilon \leq \boldsymbol{\beta}_1^\top \mathbf{x} < \boldsymbol{\beta}_2^\top \mathbf{x}^2 - 0.15 + 0.25\epsilon; \\ 3, & \text{otherwise,} \end{cases}$$

In Figure 4 we provide boxplots of average canonical correlations for Model 1 and Model 5. The data was simulated for $n = 200$ observations. Unlike in the binary case, we can now see that the results for SIR+SAVE approach the results for SIR with decreasing α. Once again the results approach those for SAVE as α increases. Overall the best performers for both models were SIR+SAVE with α approximately 0.5 and SAVE|SIR which provided similar results.

4.2. Choosing K and the number of iterations

So far we have dealt with known K. However, in practice K is usually not known and needs to be estimated. Tests for K have been developed for the various methods discussed in this paper. For example, Li ([16, 18]) introduced conservative tests for SIR and PHD respectively whilst Cook [5] and Bentler

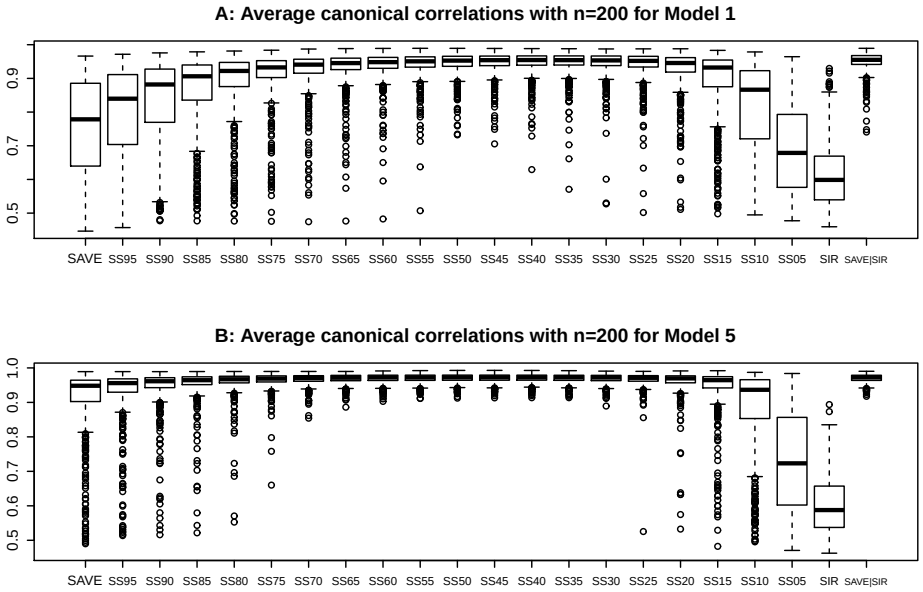

FIGURE 4. *Boxplots of the average canonical correlations for 500 iterations of data simulated from Model 1 (top) and Model 5 (bottom) for* $n = 200$. *For the Model 1 results,* $H = 5$. *The methods compared are SIR, SAVE, SAVE|SIR and SIR+SAVE with values of* $\alpha = 0.95$ *down to* $\alpha = 0.05$ *in increments of* 0.05. *SS95 refers to SIR+SAVE with* $\alpha = 0.95$, *SS90 is SIR+SAVE with* $\alpha = 0.9$, *etc.*

and Xie [1] provided further suggestions for improved tests for PHD. Zhu et al. [26] provide tests for CUME, CUVE and CUDR. Other methods exist, including those given by Bura and Cook [3] and Ferre [10].

When it comes to the number of directions to choose from the first method (i.e. OLS, SIR or CUME for the approaches we have considered here) then either:

(i) The number of initial directions may be already determined. For example, naturally OLS can only return one initial direction and SIR is restricted to returning, at most, $H - 1$ (the number of slices minus one) which is most limiting when Y is discrete where H is the number of uniquely possible responses.
(ii) A test for K or visual inspection of the eigenvalues from the eigen-decomposition step in the algorithm can be used for guidance.

For the second situation above, many tests for K are biased towards the expected value of K (i.e. the rank of the matrix in Step 1 of the methods discussed earlier) rather than the true dimension K of the underlying model. For example, we simulated 1000 trials of Model 1 for $n = 400$ and $p = 10$ and then used the test from Li [16] for SIR. The test selected $K = 1$ 95.5% of the time and when this happens the choice of one direction from SIR follows naturally.

Complementary to this, such estimates for K can be used simultaneously to both determine the number of directions to select for the first method and also the number of directions to choose from the second. For example, Zhu et al. [26] consider various models with true dimension $K = 2$ but where, due to symmetrical dependence, CUME (and also SIR) are expected to find only one direction. For the majority of their models (four of six) and 1000 simulated runs, the test for K based on the CUME eigenvalues chose the estimated K as $\widehat{K} = 1$ 100% of the time. On the other hand, the estimated K using the CUVE and CUDR eigenvalues was chosen to be $\widehat{K} = 2$ at least 80% of the time and up to 99% of the time (with the exception of one model for CUDR). Hence, the contradicting estimates for K occurring in practice would lead us to choose one direction from the first method and then focus on achieving the final element in the second. This approach can similarly be employed with SAVE|SIR and PHD|OLS.

For the iterative approaches we have chosen, we have opted for either OLS, SIR or CUME to be the first method. We have done so for two reasons. Firstly, these methods have less restrictive distributional conditions on $\mathbf{x}$ than the methods that follow next and have been shown to perform very well for a variety of models. Secondly, it is these approaches that are known to lack 'exhaustiveness' when symmetric dependency is evident. However, there is no specific reason as to why the order of application of the methods cannot be reversed. For example, OLS could follow PHD however it is less obvious how many directions one would choose for PHD and how to ensure that the chosen directions do not already include the OLS direction. Hence, the application of OLS, SIR and CUME first is an admission that these methods have missed some elements of $\mathcal{S}$ although an acknowledgement that they may have been very successful in finding a partial basis.

Finally, it may also be possible to carry out more than two iterations of dimension reduction methods. For example, we could use OLS first, followed by PHD and then finish with SIR to extract perhaps a final element that was not detected via PHD|OLS. However, thought would need to be given as to why the first two collectively may fail to find a complete basis and given that we have a preference for small K, the larger choices of K that intuitively follow from more than two iterations would still restrict the researcher in a data visualization capacity. We anticipate that combining two dimension reduction methodologies would suffice in most situations.

4.3. Examples

Example 4.1. For the first example we consider data provided by Dr. Hayley Castlehouse and analyzed as part of her PhD dissertation [4]. The data consists of soil compositions for 41 soil samples that were taken from a site in North Lincolnshire, England. We let the response be iron-oxide bound arsenic (As) in mg.kg^{-1}. The 15 predictors are depth and level of the chemical compounds PO_4^{3-}, Mg^{2+}, Cl^-, NO_3^-, NH_4^+, SO_4^{2-}, Ca^{2+}, K^+, F^-, Na^+, TIC, TOC, Fe, and Mn.

TABLE 3
Adjusted R^2 results of a fitted polynomial model of degree 2, using either the first direction (i.e. $K = 1$) or the first two directions (i.e. $K = 2$) found by each method, with $H = 6$ slices if applicable to the method

Method	Adjusted R^2 ($K = 1$)	Adjusted R^2 ($K = 2$)
SIR	0.724	0.728
SAVE	0.234	0.282
SIR+SAVE	0.226	0.262
PHD	0.026	0.280
OLS	0.808	–
CUME	0.771	0.797
CUVE	0.333	0.316
CUDR	0.272	0.556
SAVE|SIR	–	0.777
PHD|OLS	–	0.892
CUVE|CUME	–	0.789

To estimate the dimension of $\mathcal{S}$, we implement the Bayesian information criterion (BIC) recommended by Zhu et al. [26], which is based on the eigenvalues returned by the dimension reduction method utilized. When run on the CUME eigenvalues, we estimate $K = 1$. By contrast, when we use the CUDR eigenvalues, we estimate that $K = 2$. Given the contradiction between these two estimates, we will look at including some iterative approaches in our analysis.

The eleven dimension reduction methods which are compared using this data are contained in Table 3. Whilst OLS, PHD and the cumulative slicing procedures do not require the choice of a slicing parameter, a value of $H = 6$ has been chosen for the other methods. There are seven observations in each slice with the exception of the fourth slice, which contains six observations. For SIR+SAVE, a value of $\alpha = 0.5$ has been chosen, which was the value recommended by Zhu et al. in [28].

To assess the quality of the directions found, a polynomial model of degree two has been fit using multiple linear regression least squares to the data. Table 3 shows the adjusted R^2 values where either the first direction (i.e. $K = 1$) or the first two directions (i.e. $K = 2$) found by each method have been used. For all methods except CUVE, the results where $K = 2$ were an improvement when compared to those where $K = 1$. However, the improvement for SIR was almost negligible, as its adjusted R^2 value only increased from 0.724 to 0.728. This indicates that SIR found all of its useful information in the first direction, so that when it was combined with SAVE to find a second direction, any useful information that SAVE was able to find resulted in a much better adjusted R^2 value for SAVE|SIR of 0.777. A similar phenomenon has occurred with PHD|OLS. OLS on its own achieved an adjusted R^2 value of 0.808, however when combined with PHD, the extra information found by PHD was enough to improve the adjusted R^2 value to 0.892. In contrast with PHD|OLS's adjusted R^2 value of 0.892 when $K = 2$, PHD's result, when used as a standalone method, was 0.280. Similarly, SAVE's adjusted R^2 value was 0.282. These results show the value of the iterative approach to dimension reduction because the performance of PHD and SAVE has been greatly improved by conditioning on the information

found by OLS and SIR respectively. SIR+SAVE's result of adjusted R^2=0.262 (for $K = 2$) indicates that in this instance, the information found by SIR was overwhelmed by that of SAVE when the two matrices were added together with $\alpha = 0.5$. Overall, the best result was that of PHD|OLS when $K = 2$ (adjusted $R^2 = 0.892$).

Example 4.2. In the next example, we consider the Pen Digit data analysed by Zhu & Hastie [29] and Sheather et al. [23]. The Pen Digit database contains multiple samples from 44 different writers of handwritten digits from 0 to 9. There are 16 attributes (or predictors) for each digit. Whilst there is both a training dataset and a learning dataset, we focus on the training dataset, as did Sheather et al. We also focus on the 0s and 8s, since the estimated sufficient summary plots (ESSP) indicate that these two digits cannot be separated using just one direction (see Figure 5). This results in $n = 1,499$ observations, $p = 16$ predictors and a binary response.

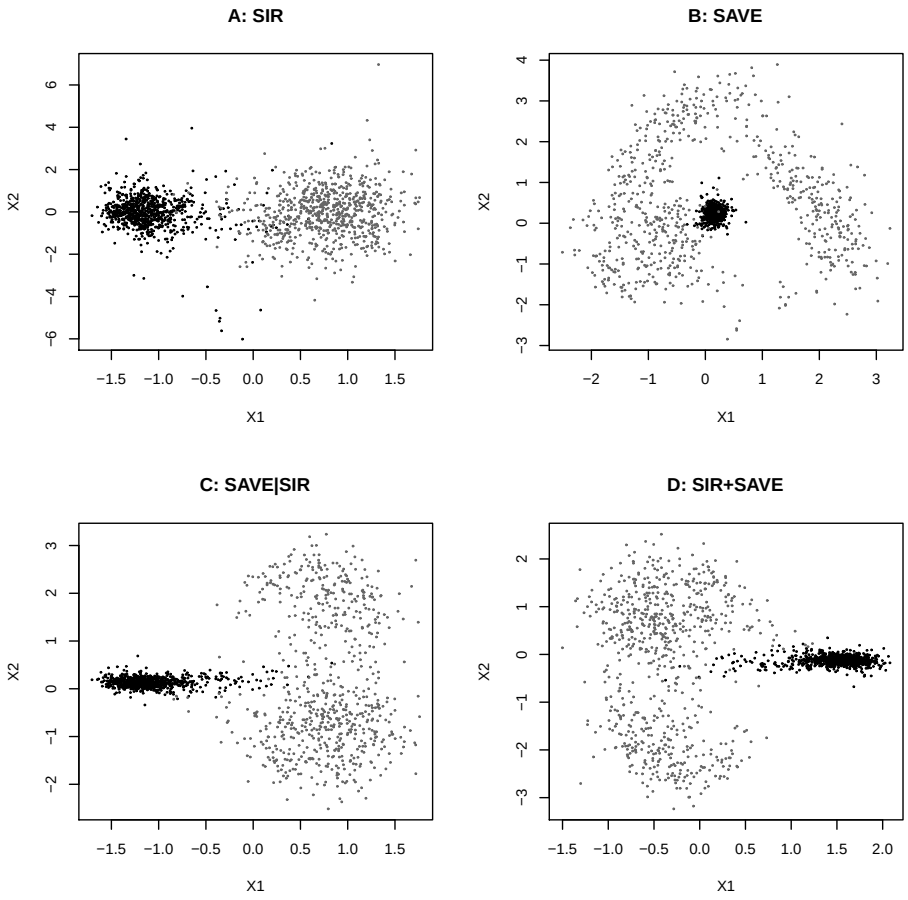

FIGURE 5. *ESSPs of the Pen Digit data for digits '0' (black) and '8' (grey).*

Due to the discrete response of the Pen Digit data, which leads to natural slicing of the data, the four methods we compare are SAVE, SIR, SAVE|SIR and SIR+SAVE with $\alpha = 0.5$. The first method of assessment we use is the ESSP plot for each method shown in Figure 5. This figure shows that SIR is able to separate the groups using the first direction, but the boundary of separation slightly overlaps. SIR's second direction, as expected, provides no further useful information since the SIR matrix is of rank one and information found in a second direction would be by chance only. The second directions of SAVE|SIR and SIR+SAVE provide useful information, as they show greater variability amongst the 8's, making separation easier. SAVE also separates well, and clearly needs two directions to do so.

The second method of assessment we use involves the use of cross validation [13] and Support Vector Machines (SVM), introduced by Karatzoglou et al. [12]. We again compare SAVE, SIR, SAVE|SIR and SIR+SAVE with $\alpha = 0.5$. Using cross validation, we use random sampling to divide the data into a training set and a testing set. Based on the directions found by the respective methods using observations in the training set, we predict the response for each observation in the testing set using a two-dimensional model via the use of SVM. To create this model, the R package e1071 [9] was used and a radial kernel was chosen. Thus, a classification rate can be found by calculating the percentage of correct predictions. This process was repeated 500 times so that average classification rates, along with standard deviations, could be found. In order to test the performance of each dimension reduction method relative to sample size, we let the number of observations in the training set take the values $n = 200$, $n = 100$, $n = 60$ and $n = 40$.

Figure 6 shows the result of this analysis. For all chosen values of n, we can see that SAVE|SIR and SIR+SAVE produce the best classification results. This supports the results of the binary response simulation in Section 4.1, and SAVE|SIR has the advantage because, unlike SIR+SAVE, it does not require the choice of the extra parameter α. In general, SIR produces the worst results because it is only able to find one direction. However, we note that whilst SAVE's average classification rate is higher than SIR's, its variability is much higher when $n = 60$ and $n = 40$, suggesting it is more volatile than SIR for this model.

5. Conclusions

In this article, we introduced a new way of combining existing dimension reduction methods. Whilst it is known that methods such as SIR, OLS and CUME are restricted in some cases, we have shown that the information found by these methods for part of $\mathcal{S}$ can still be extremely useful. We have compared this new iterative method with existing stand-alone dimension reduction methods using both simulated and real-world data with either a continuous or discrete response, and found that the method we introduce is capable of producing superior results. We also compared the new method against the combination method recommended by Zhu et al. [28], SIR+SAVE and found the results of the new

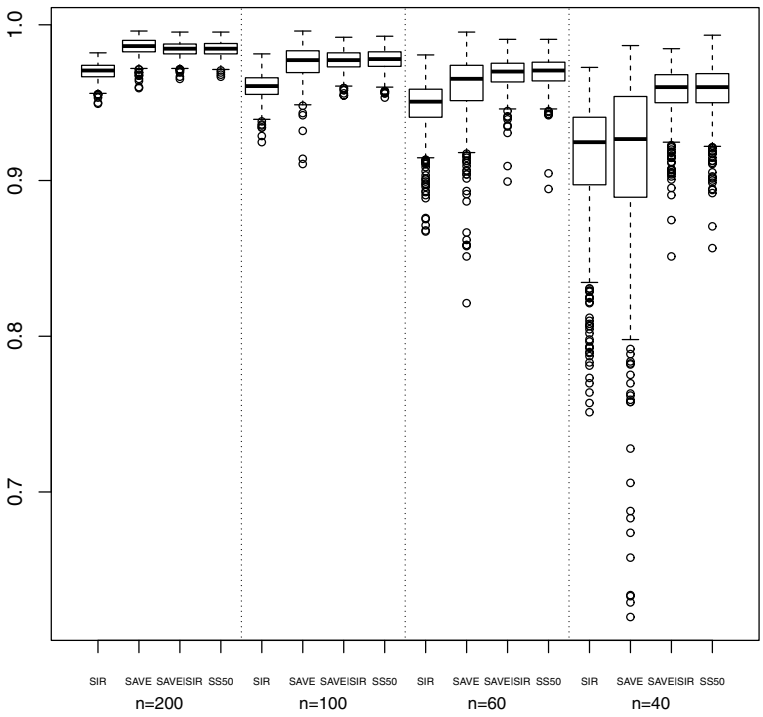

FIGURE 6. *Boxplots of classification rates for 500 random samples of the Pen Digit data for four choices of n (200, 100, 60 and 40). Methods compared are SIR, SAVE, SAVE|SIR and SIR+SAVE with* $\alpha = 0.5$ *(SS50).*

method to be at least as good as SIR+SAVE but without the added complication of having to choose the parameter α. This does not mean that our approach should necessarily be preferred over the existing combinations, but rather that it provides another alternative that may sometimes provide improved results.

Acknowledgements

The authors wish to thank the editor and two referees, whose valuable comments led to an improved version of this article. We would also like to thank Professor Li-Xing Zhu and Dr. Li-Ping Zhu, who kindly provided code relating to some of the cumulative slicing approaches. Thank you also to Dr. Hayley Castlehouse for the use of the soil data that she provided.

References

[1] BENTLER, P. M., & XIE, J. 2000. Corrections to test statistics in principal hessian directions. *Statist. Probab. Lett.*, **47**, 381–389.

[2] Brillinger, D. R. 1977. The identification of a particular nonlinear time series system. *Biometrika*, **64**, 509–515. MR0483236
[3] Bura, E., & Cook, R.D. 2001. Extending sliced inverse regression. *J. Amer. Statist. Assoc.*, **96**, 996–1003. MR1946367
[4] Castlehouse, H. 2008. *The biogeochemical controls on arsenic mobilisation in a geogenic arsenic rich soil*. PhD dissertation, University of Sheffield.
[5] Cook, R. D. 1998a. Principal Hessian directions revisited. *J. Amer. Statist. Assoc.*, **93**, 84–100. With comments by Ker-Chau Li and a rejoinder by the author. MR1614584
[6] Cook, R. D. 1998b. *Regression graphics*. New York: John Wiley & Sons Inc. Ideas for studying regressions through graphics. MR1645673
[7] Cook, R. D., & Li, B. 2002. Dimension reduction for conditional mean in regression. *Ann. Statist.*, **30**, 455–474. MR1902895
[8] Cook, R.D., & Weisberg, S. 1991. Sliced inverse regression for dimension reduction: Comment. *J. Amer. Statist. Assoc.*, **86**, 328–332. MR1137117
[9] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. 2010. *e1071: Misc functions of the department of statistics (e1071), tu wien*. R package version 1.5-24.
[10] Ferre, L. 1998. Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.*, **93**, 132–140. MR1614604
[11] Hall, P., & Li, K.-C. 1993. On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.*, **21**, 867–889.
[12] Karatzoglou, A., Meyer, D., & Kurt, H. 2005. Support vector machines in R. Department of Statistics and Mathematics, Wirtschaftsuniversität Wien: ePub, http://epub.wu-wien.ac.at.
[13] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Pages 1137–1145 of: IJCAI*.
[14] Li, B., & Wang, S. L. 2007. On directional regression for dimension reduction. *J. Amer. Statist. Assoc.*, **102**, 997–1008. MR2354409
[15] Li, B., Zha, H. Y., & Chiaromonte, F. 2005. Contour regression: a general approach to dimension reduction. *Ann. Statist.*, **33**, 1580–1616. MR2166556
[16] Li, K.-C. 1991a. Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, **86**, 316–342. MR1137117
[17] Li, K.-C. 1991b. Rejoinder for discussions on sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.*, **86**, 337–342. MR1137117
[18] Li, K.-C. 1992. On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.*, **87**, 1025–1039. MR1209564
[19] Li, K.-C., & Duan, N. 1989. Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.

[20] Lue, H.-H., Chen, C.-H, & Chang, W.-H. 2011. Dimension reduction in survival regressions with censored data via an imputed spline approach. *Biometrical journal*, **53**, 426–443.
[21] Prendergast, L. A. 2005. Influence functions for sliced inverse regression. *Scand. J. Statist.*, **32**, 385–404. MR2204626
[22] Prendergast, L. A., & Smith, J. A. 2010. Influence functions for dimension reduction methods: An example influence study of principal hessian direction analysis. *Scand. J. Statist.*, **37**, 588–611.
[23] Sheather, S.J., McKean, J.W., & Crimin, K. 2008. Sliced mean variance-covariance inverse regression. *Comput. Statist. Data Anal.*, **52**, 1908–1927. MR2418480
[24] Stein, C. M. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135–1151. MR0630098
[25] Ye, Z., & Weiss, R.E. 2003. Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.*, **98**, 968–979. MR2041485
[26] Zhu, L. P., Zhu, L. X., & Feng, Z. H. 2010a. Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.*, **105**, 1455–1466. MR2796563
[27] Zhu, L. P., Wang, T., Zhu, L. X., & Ferre, L. 2010b. Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295–304. MR2650739
[28] Zhu, L. X., Ohtaki, M., & Li, Y. 2007. On hybrid methods of inverse regression-based algorithms. *Comput. Statist. Data Anal.*, **51**, 2621–2635. MR2338992
[29] Zhu, M., & Hastie, T.J. 2003. Feature extraction for nonparametric discriminant analysis. *J. Comput. Graph. Statist.*, **12**, 101–120. MR1965211