Using Sliced Inverse Mean Difference for Sufficient Dimension Reduction

Andreas Artemiou[a,*], Lipu Tian[b]

[a]*School of Mathematics, Cardiff University*
[b]*Department of Mathematical Sciences, Michigan Technological University*

Abstract

We present two different algorithms for sufficient dimension reduction based on the difference between inverse means. We discuss the theoretical properties and demonstrate the computational advantages over SIR (Li, 1991) and CUME (Zhu, Zhu and Feng, 2010).

Keywords: Inverse moments; Inverse regression; Sequential tests; Sufficient Dimension Reduction

1. Introduction

Sufficient dimension reduction methods aim to reduce the dimension of a regression problem via supervised dimension reduction without losing information of the conditional distribution of $Y|\boldsymbol{X}$ where Y is the response and $\boldsymbol{X}$ is a p dimensional predictor. Thus, one estimates a $p \times d$ ($d \leq p$) matrix $\boldsymbol{\beta}$ under the model

$$Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X} \tag{1}$$

If $d < p$ dimension reduction is achieved. For each $\boldsymbol{\beta}$ satisfying model (1) we define the Dimension Reduction Subspace (DRS), denoted by $\mathcal{S}(\boldsymbol{\beta})$, to be the space spanned by the column vectors of $\boldsymbol{\beta}$. The intersection of all DRSs, if it is a DRS itself, it is the minimum dimension reduction subspace, it is unique, and it is known as the Central Subspace (CS), denoted by $\mathcal{S}_{Y|\boldsymbol{X}}$ (see Cook - 1998a). The ultimate goal is to estimate accurately the matrix $\boldsymbol{\beta}$ whose column space span the CS. Conditions of existence of the CS are given in Cook (1998a) and

*Corresponding Author

Yin, Li and Cook (2008). These conditions are mild and throughout this paper we assume the existence of CS.

Li (1991) proposed Sliced Inverse Regression (SIR), the first method introduced in the sufficient dimension reduction framework. A number of methods followed, for instance Sliced Average Variance Estimator (SAVE) by Cook and Weisberg (1991), principal Hessian directions (pHd) by Li (1992), Contour Regression (CR) by Li, Zha, Chiaromonte (2005), Directional Regression (DR) by Li and Wang (2007) among others. Each of these methods performs linear sufficient dimension reduction with its own advantages and limitations.

SIR, SAVE and DR use the idea of slicing the response variable to perform dimension reduction. The number of slices is a tuning parameter of these algorithms. SAVE and DR were shown to have performance highly influenced by the number of slices. More recently, Zhu, Zhu and Feng (2010) developed three new algorithms based on the aforementioned three methods, using cumulative slicing, and named the algorithms Cumulative Mean Estimation (CUME), Cumulative Variance Estimation (CUVE) and Cumulative Directional Regression (CUDR). This created the cumulative idea where there was no need to tune for the number of slices as one starts from the first point (the smaller value of the response) and add one point at each iteration of the algorithm. There are two concerns with CUME. First, as n increases the number of cutoff points increase significantly, and in today's world with massive datasets being the norm in many sciences, this can cause computational problems. Second, the cumulative nature of the algorithm makes it inappropriate for problems with categorical response where there is no natural ordering of the categories.

Here we propose a new approach which uses the idea of slicing the response but use the difference between inverse means of two slices to achieve dimension reduction. We propose two different algorithms to estimate the CS. The first algorithm is equivalent to CUME in theory but is faster computationally and the second algorithm can handle categorical responses better. These two algorithms are based on the "left vs right" (LVR) and "one vs another" (OVA) algorithms presented in Li, Artemiou and Li (2011).

The paper is constructed as follows. In Section 2, the idea is introduced with more details. In Section 3, we give an estimation algorithm, the asymptotic properties and discuss dimension determination of the CS. Numerical studies

follow on section 4. Finally, we conclude with a discussion. All the proofs are in a supplementary file.

2. Inverse Mean Difference approach

Let Y be a response variable in a regression problem and $\boldsymbol{X}$ be a p dimensional predictor vector, where for simplicity we assume $E(\boldsymbol{X}) = 0$. Let n be the number of observations in our regression problem and H the number of slices.

Sliced inverse regression (SIR) by Li (1991) uses the idea of the inverse first moment to estimate the CS. More specifically a spectral decomposition of $(\boldsymbol{\Sigma})^{-1}\text{var}(E(\boldsymbol{X}|Y))$ is used where $\boldsymbol{\Sigma} = \text{var}(\boldsymbol{X})$. More recently Zhu, Zhu and Feng (2010) proposed the so-called Cumulative Mean (CUME) estimation of $\mathcal{S}_{Y|\boldsymbol{X}}$. They proposed to estimate $\mathcal{S}_{Y|\boldsymbol{X}}$ using a spectral decomposition of $(\boldsymbol{\Sigma})^{-1}\text{var}(E(\boldsymbol{X}I(Y \leq k)))$ where k is any value satisfying $y_{(1)} \leq k \leq y_{(n)}$ and $y_{(i)}$ is the i^{th} ordered value of the response.

Here we propose to estimate CS using the difference of the means of two disjoint set of points. Let Ω be the support of Y and let A_1 and A_2 be two disjoint subsets of Ω. Then using $I(\cdot)$ as the indicator function the response variable Y can be discretized using:

$$\tilde{Y} = I(Y \in A_1) - I(Y \in A_2). \tag{2}$$

We propose to estimate the CS using the difference of the means between the points that belong to the two sets A_1, A_2. We denote this as:

$$m_d = E(\boldsymbol{X}I(\tilde{Y} = 1)) - E(\boldsymbol{X}I(\tilde{Y} = -1)). \tag{3}$$

It is pretty straightforward to prove the following result. The assumption $E(\boldsymbol{X}|\boldsymbol{\beta}^{\intercal}\boldsymbol{X}) = \boldsymbol{P}^{\intercal}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}$ is a very common assumption in the SDR literature and it holds if the predictors are elliptically distributed.

Theorem 1. *If* $E(\boldsymbol{X}|\boldsymbol{\beta}^{\intercal}\boldsymbol{X}) = \boldsymbol{P}^{\intercal}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma})\boldsymbol{X}$ *where* $\boldsymbol{P}^{\intercal}_{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \boldsymbol{\beta}(\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}$. *Then* $m_d \in \mathcal{S}_{Y|\boldsymbol{X}}$.

We use two different algorithms based on the way the sets A_1, A_2 are defined in (2). The first approach is called "left vs right" (LVR). If we divide the dataset into H slices, this approach uses the $H - 1$ cutoff points between the slices, denoted as $q_r, r = 1, \ldots, H - 1$. Using the cutoff point q_r we define $\tilde{Y}^r_{LVR} =$

$I(Y > q_r) - I(Y \leq q_r)$. This means that A_1 contains the points with response greater than the cutoff point and A_2 contains the points less than or equal to the cutoff point. Under this method equation (3) becomes $m_{\text{LVR}}(q_r) = E(\boldsymbol{X}I(Y > q_r)) - E(\boldsymbol{X}I(Y \leq q_r))$. One can show that this approach is equivalent to CUME. Since we assume $E(\boldsymbol{X}) = 0$ it follows that $E(\boldsymbol{X}I(Y > q_r)) = -E(\boldsymbol{X}I(Y \leq q_r))$ which implies that $m_{\text{LVR}}(q_r) = -2E(\boldsymbol{X}I(Y \leq q_r))$ which is twice the CUME estimator. Although this is theoretically equivalent to CUME computationally it is faster, especially when n gets very large, as instead of using n cutoff points as CUME does, it uses only $H - 1$ which is usually much less than n.

The second algorithm is called "one vs another" (OVA). Dividing the dataset into H slices, we select a pair of slices $(i, j), i > j, i, j = 1, \ldots, H$. Under this method equation (3) becomes $m_{\text{OVA}}(i, j) = E(\boldsymbol{X}I(Y \in H_i)) - E(\boldsymbol{X}I(Y \in H_j))$ where H_i denotes the i^{th} slice. Using this method there are $\binom{H}{2}$ pairs and there is no sense of ordering as in the LVR method. Therefore this method might be more suitable for categorical responses where no ordering exists. Interestingly, in the special case that all slices have an equal number of observations this approach is equivalent to SIR.

We call this method the Slice Inverse Mean Difference (SIMD) method and to distinguish between the two algorithms when necessary we will use the subscripts LVR and OVA.

3. Statistical Inference

In this section we first outline the algorithm for sample estimation for both methods; we then provide some asymptotic results and finally develop sequential tests for estimating the dimension of the CS only for LVR.

3.1. Sample estimation

Having a set of n observations $(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n$ the following steps are used to estimate $\mathcal{S}_{Y|\boldsymbol{X}}$:

1. Let $\boldsymbol{Z}_i$ be the standardized version of $\boldsymbol{X}_i$, that is set $\boldsymbol{Z}_i = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\boldsymbol{X}_i - \bar{X})$ where $\bar{X}$ the mean of the $\boldsymbol{X}_i$'s and $\hat{\boldsymbol{\Sigma}}$ the estimate of the var$(\boldsymbol{X})$.
2. Divide the range of the response variable into H slices. Let $q_r, r = 1, \ldots, H - 1$ be the dividing points between the slices.

3a. (LVR) For each $q_r, r = 1, \ldots, H-1$ define the discretized response variable $\tilde{Y}_i^r = I(Y_i > q_r) - I(Y_i \leq q_r)$ and calculate

$$\hat{m}^Z_{\text{LVR}}(q_r) = \frac{1}{n_1^r}\frac{n_1^r}{n}\sum_{i=1}^{n} \boldsymbol{z}_i I(\tilde{Y}_i^r = 1) - \frac{1}{n_{-1}^r}\frac{n_{-1}^r}{n}\sum_{i=1}^{n} \boldsymbol{z}_i I(\tilde{Y}_i^r = -1) \quad (4)$$

where $n_j^r, j = -1, 1$ denotes the number of points with discrete value $Y_i^r = j$ at dividing point q_r.

3b. (OVA) For each pair (r, s) satisfying $1 \leq r < s \leq H$ define the discretized response $\tilde{Y}_i^{rs} = I(A_{s-1} < Y_i \leq A_s) - I(A_{r-1} < Y_i \leq A_r)$ where A_i defines the i^{th} ordered slice of the responses and calculate

$$\hat{m}^Z_{\text{OVA}}(r, s) = \frac{1}{n_1^{rs}}\frac{n_1^{rs}}{n}\sum_{i=1}^{n} \boldsymbol{z}_i I(\tilde{Y}_i^{rs} = 1) - \frac{1}{n_{-1}^{rs}}\frac{n_{-1}^{rs}}{n}\sum_{i=1}^{n} \boldsymbol{z}_i I(\tilde{Y}_i^{rs} = -1) \quad (5)$$

where $n_j^{r,s}, j = -1, 1$ denotes the number of points with discrete value $Y_i^{rs} = j$ for the pair of slices (r, s).

4a. (LVR) Construct the $p \times (H-1)$ matrix $\hat{\boldsymbol{\Gamma}}_{\text{LVR}}$ where each column is one of vectors $\hat{m}^Z_{\text{LVR}}(q_r)$ and use construct

$$\hat{\boldsymbol{V}}_{\text{LVR}} = \hat{\boldsymbol{\Gamma}}_{\text{LVR}}\hat{\boldsymbol{\Gamma}}_{\text{LVR}}^{\mathsf{T}} = \sum_{r=1}^{H-1} \hat{m}^Z_{\text{LVR}}(q_r)(\hat{m}^Z_{\text{LVR}}(q_r))^{\mathsf{T}} \quad (6)$$

4b. (OVA) Construct the $p \times \binom{H}{2}$ matrix $\hat{\boldsymbol{\Gamma}}_{\text{OVA}}$ where each column is one of vectors $\hat{m}^Z_{\text{OVA}}(q_r)$ and construct

$$\hat{\boldsymbol{V}}_{\text{OVA}} = \hat{\boldsymbol{\Gamma}}_{\text{OVA}}\hat{\boldsymbol{\Gamma}}_{\text{OVA}}^{\mathsf{T}} = \sum_{1 \leq r < s \leq H} \hat{m}^Z_{\text{OVA}}(r, s)(\hat{m}^Z_{\text{OVA}}(r, s))^{\mathsf{T}} \quad (7)$$

5. Find the d eigenvectors $\hat{u}_1, \ldots, \hat{u}_d$ corresponding to the d nonzero eigenvalues of $\hat{\boldsymbol{V}}$. Let $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_d)$ and use the subspace spanned by $\boldsymbol{\Sigma}^{-\frac{1}{2}}\hat{\boldsymbol{u}}$ to estimate $\mathcal{S}_{Y|\boldsymbol{X}}$.

3.2. Asymptotic normality of $\boldsymbol{\Gamma}_{\text{LVR}}$

In this section we derive the asymptotic distribution for $\hat{\boldsymbol{\Gamma}}_{\text{LVR}}$. Let $\boldsymbol{V}_{\text{LVR}}$ be the population version of matrix $\hat{\boldsymbol{V}}_{\text{LVR}}$ in equation (6)

$$\boldsymbol{V}_{\text{LVR}} = \boldsymbol{\Gamma}_{\text{LVR}}\boldsymbol{\Gamma}_{\text{LVR}}^{\mathsf{T}} = \sum_{r=1}^{H-1} m^Z_{\text{LVR}}(q_r)(m^Z_{\text{LVR}}(q_r))^{\mathsf{T}}. \quad (8)$$

where $\mathbf{\Gamma}_{\text{LVR}} = (m^Z_{\text{LVR}}(q_1), \ldots, m^Z_{\text{LVR}}(q_{H-1}))$. Note that this algorithm is based on all the points to calculate each $m^Z_{\text{LVR}}(q_r)$ which means $m^Z_{\text{LVR}}(q_i)$ and $m^Z_{\text{LVR}}(q_j)$ for any pair $(i, j), i, j = 1, \ldots, H-1$ are not independent. Therefore we rewrite them using the intraslice means which are independent. So:

$$\begin{aligned} m^Z_{\text{LVR}}(q_r) =& E(\boldsymbol{Z}I(Y > q_r)) - E(\boldsymbol{Z}I(Y \leq q_r)) = \sum_{i=r+1}^{H} E(\boldsymbol{Z}I(Y \in A_i)) - \sum_{i=1}^{r} E(\boldsymbol{Z}I(Y \in A_i)) \\ &= \sum_{i=r+1}^{H} p_i E(\boldsymbol{Z}|Y \in A_i)) - \sum_{i=1}^{r} p_i E(\boldsymbol{Z}|Y \in A_i)) \end{aligned} \tag{9}$$

where A_i denotes the i^{th} slice and p_i the proportion of points in slice A_i, $i = 1, \ldots, H-1$.

We now define $\tilde{\boldsymbol{Z}}_n = \sqrt{n}(\hat{\boldsymbol{Z}}_n - \boldsymbol{B})$. Note that $\boldsymbol{B}$ is a $p \times H$ matrix where each column is $p_i E(\boldsymbol{Z}|Y \in A_i)$, $i = 1, \ldots, H$ and $\hat{\boldsymbol{Z}}_n$ is the sample version of $\boldsymbol{B}$. Using the multivariate central limit theorem and the multivariate version of Slutsky's theorem one can prove the following result:

Lemma 1. *Let $\mathbf{\Sigma}_{z|s} = \text{cov}(\boldsymbol{Z}|s)$ and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Then $\text{vec}(\tilde{\boldsymbol{Z}}_n) \xrightarrow{\mathcal{D}} N_{pH}(0, \mathbf{\Delta})$ where $\mathbf{\Delta}$ is a $pH \times pH$ matrix which is an $H \times H$ array of $p \times p$ matrices $\mathbf{\Delta}_{ts}$ where for $t = s$ we have $\mathbf{\Delta}_{ss} = \boldsymbol{I}_p p_s^2 + (1 - 2p_s)\mathbf{\Sigma}_{z|s}$ and for $t \neq s$ we have $\mathbf{\Delta}_{ts} = p_t p_s(\boldsymbol{I}_p - \mathbf{\Sigma}_{z|s} - \mathbf{\Sigma}_{z|t})$.*

The proof is similar to a result in Bura and Cook (2001) and is omitted.

The above result is used together with the Delta method to prove the following result which gives the asymptotic distribution of $\hat{\mathbf{\Gamma}}_{\text{LVR}}$.

Theorem 2.

$$\sqrt{n}\text{vec}(\hat{\mathbf{\Gamma}}_{\text{LVR}} - \mathbf{\Gamma}_{\text{LVR}}) \xrightarrow{\mathcal{D}} N_{p(H-1)}(0, \boldsymbol{W}\mathbf{\Delta}\boldsymbol{W}^{\intercal})$$

where $\boldsymbol{W}$ is a $p(H-1) \times pH$ matrix which is an $(H-1) \times H$ array of $p \times p$ positive or negative identity matrices. Denoting by $\boldsymbol{W}_{ij}$ the element at the i^{th} row and j^{th} column of the array $\boldsymbol{W}$, $\boldsymbol{W}_{ij} = \boldsymbol{I}$ if $j > i$ and $\boldsymbol{W}_{ij} = -\boldsymbol{I}$ if $j \leq i$.

Similar results for $\hat{\mathbf{\Gamma}}_{\text{OVA}}$ are in the supplementary file.

3.3. Dimension determination through sequential tests

In this section we develop sequential tests to determine d, the dimension of $\mathcal{S}_{Y|\boldsymbol{X}}$. Sequential tests are frequently used in the literature. For SIR, see Li

(1991), Schott (1994), Velilla (1998), Ferré (1998), Bura and Cook (2001); for SAVE see Shao, Cook and Weisberg (2007); for pHd see Cook (1998b); for DR see Li and Wang (2007). Bura and Yang (2011) developed a unifying approach.

The developments here follow the results of Bura and Yang (2011). For the rest of the section the variance of the asymptotic distribution of $\hat{\boldsymbol{\Gamma}}_{\text{LVR}}$ is denoted by $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Gamma}}} = \boldsymbol{W}^{\intercal}\boldsymbol{\Delta}\boldsymbol{W}$ and we avoid the use of the LVR subscript throught the section.

Assuming $\text{rank}(\boldsymbol{\Gamma}) = k = \min(p, H-1)$, then the singular value decomposition of matrix $\boldsymbol{\Gamma}$ is given by:

$$\boldsymbol{\Gamma} = \boldsymbol{U}^{\intercal} \begin{pmatrix} \boldsymbol{D}_1 & \mathbf{0}_{k,p-k} \\ \mathbf{0}_{p-k,k} & \mathbf{0}_{p-k,p-k} \end{pmatrix} \boldsymbol{R}$$

where $\mathbf{0}_{i,j}$ is the $i \times j$ matrix with all entries equal to 0, $\boldsymbol{U}^{\intercal} = (\boldsymbol{U}_1, \boldsymbol{U}_0)$ is a $p \times p$ orthogonal matrix with the left singular vectors of $\boldsymbol{\Gamma}$ where $\boldsymbol{U}_1$ is a $p \times k$ matrix of the k left singular vectors vectors corresponding to the largest singular values and $\boldsymbol{U}_0$ is a $p \times (p-k)$ matrix, $\boldsymbol{D}_1 = \text{diag}(\lambda_1, \ldots, \lambda_k)$ is a $k \times k$ matrix where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$ and $\boldsymbol{R}^{\intercal} = (\boldsymbol{R}_1, \boldsymbol{R}_0)$ is an $(H-1) \times (H-1)$ matrix with the right singular values of $\boldsymbol{\Gamma}$ where $\boldsymbol{R}_1$ is a $(H-1) \times k$ matrix having the k left singular vectors vectors corresponding to the largest singular values and $\boldsymbol{R}_0$ is a $(H-1) \times (H-1-k)$ matrix.

Similarly, we can have the singular value decomposition of matrix $\hat{\boldsymbol{\Gamma}}$, which is given by:

$$\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{U}}^{\intercal} \begin{pmatrix} \hat{\boldsymbol{D}}_1 & \mathbf{0}_{k,p-k} \\ \mathbf{0}_{p-k,k} & \hat{\boldsymbol{D}}_0 \end{pmatrix} \hat{\boldsymbol{R}}$$

where we use the sample estimates of matrices $\boldsymbol{U}$, $\boldsymbol{R}$ and $\boldsymbol{D}_1$. $\hat{\boldsymbol{D}}_0$ is the estimator of $\mathbf{0}_{p-k,p-k}$.

Assume now we have the following sequential tests $H_0 : \text{rank}(\boldsymbol{\Gamma}) = k$ vs $H_A : \text{rank}(\boldsymbol{\Gamma}) > k$, $k = 0, \ldots, p$. Starting with $k = 0$, we test the above hypothesis. If the null is rejected we repeat the test increasing the value of k by 1. The smallest value of k the null hypothesis is not rejected is assumed to be the true rank of matrix $\boldsymbol{\Gamma}$.

Define now the following test statistic: $T_1(k) = n\text{vec}(\hat{\boldsymbol{D}}_0)^{\intercal}\text{vec}(\hat{\boldsymbol{D}}_0) = \sum_{i=k+1}^{\min(p,H-1)} \hat{\lambda}_i^2$ where $\hat{\lambda}_i$'s the singular values of $\hat{\boldsymbol{\Gamma}}$. Then by a direct application of Theorem 1 in Bura and Yang (2011) we have the following corollary:

Corollary 1. *Assume the assumptions of Theorem 2 hold and* $\text{rank}(\boldsymbol{\Gamma}) = k$. *Then* $T_1(k) \xrightarrow{d} \sum_{i=1}^{s} w_i K_i$ *where* $K_i \sim \chi^2_i$ *and* $w_1 \geq \ldots \geq w_s$ *are the ordered eigenvalues of* $\boldsymbol{Q} = (\boldsymbol{R}_0^\top \otimes \boldsymbol{U}_0^\top)\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Gamma}}}(\boldsymbol{R}_0 \otimes \boldsymbol{U}_0)$ *and* $s = \min(\text{rank}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Gamma}}}), (p-k)(H-1-k))$

Define a second test statistic $T_2(k) = n\text{vec}\left(\hat{\boldsymbol{D}}_0^\top\right)\hat{\boldsymbol{Q}}^+\text{vec}\left(\hat{\boldsymbol{D}}_0^\top\right)$ where $\hat{\boldsymbol{Q}}^+$ is the estimator of the Moore-Penrose inverse of matrix $\boldsymbol{Q}$ in Corollary 1. Then by direct application of Theorem 2 in Bura and Yang (2011) we have the following corollary:

Corollary 2. *Assuming the assumptions of Theorem 2 hold and* $\text{rank}(\boldsymbol{\Gamma}) = k$. *Then* $T_2(k) \xrightarrow{d} \chi^2_s$ *where* $s = \min(\text{rank}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Gamma}}}), (p-k)(H-1-k))$.

For $T_1(k)$, Bentler and Xie (2000) proposed two approximations. The first is the scaled version, $T_{sc}(k) = \frac{T_1(k)}{c} \sim \chi^2_s$ where $c = \sum_{i=1}^{s} \frac{w_i}{s}$ and w_i and s as defined in Corollary 1. The second is the adjusted version, $T_{adj}(k) = \frac{T_1(k)}{a} \sim \chi^2_b$ where $a = \frac{\sum_{i=1}^{s} w_i^2}{\sum_{i=1}^{s} w_i}$ and $b = \frac{\left(\sum_{i=1}^{s} w_i\right)^2}{\sum_{i=1}^{s} w_i^2}$.

Similar results hold for the OVA algorithm by replacing $H-1$ with $\binom{H}{2}$.

4. Numerical Studies

In this section we present numerical studies to demonstrate the advantages of the new algorithms. To compare performance between different methods we use the trace correlation defined by Ferré (1998)

$$r(K) = \frac{\text{trace}\boldsymbol{P}_{\mathcal{S}}\boldsymbol{P}_{\hat{\mathcal{S}}}}{K} \tag{10}$$

which uses the trace of a matrix where $\mathcal{S}$ is the true space and $\hat{\mathcal{S}}$ is the estimated space, $\boldsymbol{P}_A$ is the projection operator in the standard inner product of A, and K is the dimension of $\mathcal{S}$. $r(K)$ takes values between 0 and 1 and the closest it is to 1, the closest the true space and the estimated space are.

4.1. Performance of estimation

For our simulations we use the models I: $Y = X_1/[0.5 + (X_2 + 1)^2] + \sigma\varepsilon$ and II: $Y = X_1(X_1 + X_2 + 1) + \sigma\varepsilon$ where $\boldsymbol{X} \sim N_p(0, \boldsymbol{I}_p)$, $\varepsilon \sim N(0, 1)$ and $\sigma = 0.2$. We run 500 simulations of each experiment setup with sample size 100.

Table 1 SIMD_{LVR} and CUME perform slightly better than SIR, although CUME has smaller variability (especially for model I). We can see that as we

Table 1: Average trace correlation (with standard deviation in parenthesis) of SIR, CUME and SIMD_{LVR} with different numbers of slices H and different values for the dimension of the predictor p.

Models	p	H	SIR	CUME	SIMD_{LVR}
I	10	10	0.81 (0.097)	0.86 (0.062)	0.85 (0.067)
		20	0.77 (0.124)		0.85 (0.069)
	20	10	0.66 (0.104)	0.74 (0.073)	0.72 (0.073)
		20	0.59 (0.105)		0.71 (0.075)
	30	10	0.55 (0.093)	0.64 (0.075)	0.63 (0.070)
		20	0.49 (0.088)		0.61 (0.071)
II	10	10	0.62 (0.162)	0.69 (0.128)	0.72 (0.124)
		20	0.54 (0.162)		0.72 (0.122)
	20	10	0.42 (0.146)	0.50 (0.131)	0.53 (0.124)
		20	0.36 (0.142)		0.56 (0.134)
	30	10	0.29 (0.121)	0.37 (0.117)	0.40 (0.114)
		20	0.23 (0.108)		0.42 (0.120)

increase the number of slices, SIR lose accuracy while SIMD_{LVR} maintains the performance it has for small number of slices. Although not shown, for this set of experiments SIMD_{OVA} performs exactly as SIR.

4.2. Computation time

To demonstrate the improvement in the performance of the two algorithms we run model I, with $p = 10$ and $n = 10^2, 10^3, 10^4, 10^5$. In one set of experiments $H = 10$ and in the second set $H = n/10$. The results are summarized in Table 2. The computational time of SIMD is clearly shorter even in the case that

Table 2: Time in seconds to execute one iteration of SIMD_{LVR} and CUME for different sample sizes

n	SIMD ($H = 10$)	SIMD ($H = n/10$)	CUME
10^2	0.005	-	0.014
10^3	0.010	0.101	0.345
10^4	0.074	6.166	29.113
10^5	0.463	458.050	2169.469

Table 3: Percentage of accurate prediction of the effective dimension by SIR and SIMD_{LVR}. For SIR the unifying approach developed by Bura and Yang (2011) was used. The number of slices is 10.

		n=200		n=400		n=500	
Models	p	SIR	SIMD_{LVR}	SIR	SIMD_{LVR}	SIR	SIMD_{LVR}
I, $d=2$	10	94	100	91	100	96	100
	20	87	95	91	100	94	100
	30	75	77	93	100	87	100
II, $d=2$	10	61	85	91	100	90	100
	20	63	67	80	100	96	100
	30	47	37	79	98	86	99
III, $d=1$	10	91	93	90	94	92	91
	20	80	94	88	94	93	95
	30	66	97	89	95	89	97
IV, $d=1$	10	94	97	92	94	92	96
	20	82	95	90	99	89	95
	30	70	100	83	96	87	96

we use a huge number of slices. Although not shown we emphasize that the performance of SIMD is not affected by the different number of slices and it is very close to CUME for all sample sizes n (even when $H = 10$).

4.3. Performance for order determination

We run a simulation to compare the performance of the sequential tests developed in the previous section for the SIMD_{LVR}. We compare the performance of those tests with the tests developed for SIR in Bura and Cook (2001). We use models I, II . We also include Model III:$Y = X_1 + X_2 + \sigma\varepsilon$, Model IV:$Y = X_1/[0.5 + (X_1 + 1)^2] + \sigma\varepsilon$. The effective dimension for models III and IV is $d = 1$ and for models I and II the effective dimension is $d = 2$. The results are summarized in Table 3 where only the results for test statistic T_{sc} are presented for briefness. Under most scenarios the tests for SIMD_{LVR} work slightly better.

4.4. Categorical responses

Table 4: First direction coefficients for SIMD_{OVA} and CUME for the Iris data under order 1 (setosa=1, versicolor=2, viginica=3) and order 2 ((setosa=2, versicolor=1, viginica=3)). The last row gives the distance between the two SIMD_{OVA} vectors and the distance between the two CUME vectors

	SIMD_{OVA}		CUME	
Variables	Order 1	Order 2	Order 1	Order 2
Petal Length	-0.150	-0.150	-0.149	-0.091
Petal Width	-0.148	-0.148	-0.066	0.188
Sepal Length	0.851	0.851	0.714	0.063
Sepal Width	0.481	0.481	0.681	0.976
Distance	dist=1		dist=0.505	

We use the Iris data to demonstrate the advantage of SIMD_{OVA} with categorical responses. The dataset consists of 150 observations, 50 from each of setosa, versicolor, virginica species of iris flower. For each flower petal length and width and sepal length and width are measured. Since there is no natural ordering of the species we run the OVA algorithm and CUME using two different orderings of the species. In the first run setosa, versicolor and virginica are coded as 1, 2, 3 respectively and in the second they are coded as 2, 1, 3 respectively. Table 4 shows the first direction extracted by each method for each ordering. It is clear that for OVA there is no difference, while there is a big difference for CUME. The distance measured is based on the trace correlation in (10).

5. Discussion

In this work we use the differences of inverse means to achieve sufficient dimension reduction. We present two different algorithms to achieve this. The first algorithm, called LVR, is theoretically equivalent to CUME by Zhu, Zhu and Feng (2010) but has certain advantages. First, when the number of observations is really large it is estimating the CS faster than CUME as it uses much less cutoff points. Also if it is compared to SIR it is more robust to the number of slices. The second algorithm, called OVA, is shown to solve the issue CUME has when the response is categorical with no logical ordering between its values.

We believe similar algorithms can be developed for SAVE and DR algorithms. Since those two methods use conditional second moments we believe they require different methods treatment as one should make sure some properties of covariance matrices are not affected by using functions of two covariance matrices.

Acknowledgements

We are very grateful to an associate editor, a referee and Dr. Jonathan Gillard, whose many useful comments and suggestions greatly broadened the results of this manuscript.

References

1. Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. Bentler, P. M. and Xie, J. (2000). Corrections to test statistics in principal Hessian directions. *Statistics and Probability letters*, **47**, 381–389.
3. Bura, E. and Cook, D. R. (2001). Extending SIR: the weighted chi-square test. *Journal of the American Statistical Association*, **96**, 996–1003.
4. Bura, E. and Pfeiffer, R. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters*, **78**, 2275–2280.
5. Bura, E. and Yang, J. (2011). Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, **102**, 130–142
6. Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
7. Cook R. D. (1998b). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, **93**, 84 – 100.
8. Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *The Annals of Statistics*, **30**, 455–474.
9. Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, **86**, 316–342.

10. Ferré, L. (1998). Determining the Dimension in Sliced Inverse Regression and related methods. *Journal of the American Statistical Association*, **93**, 132–140.
11. Horton, P. and Nakai, K. (1996). A Probablistic Classification System for Predicting the Cellular Localization Sites of Proteins. *Intelligent Systems in Molecular Biology*, **4**, 109–115.
12. Li, B., Artemiou, A. and Li L. (2011). Principal Support Vector Machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, **39**, 3182–3210
13. Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997–1008.
14. Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580-1616.
15. Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316 – 342.
16. 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*, **86**, 316 – 342.
17. Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communication in Statistics - Theory and Methods*, **29**, 2141–2171
18. Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, **89**, 141–148.
19. Shao, Y. Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika*, **94**, 285–296.
20. Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association*, **93**, 1088–1098.
21. 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.
22. Zhu L. P., Zhu L. X. and Feng Z. H. (2010) Dimension Reduction in Regression through Cumulative Slicing Estimation *Journal of the American Statistical Association*, **105**, 1455-1466.