Representative approach for big data dimension reduction with binary responses

Xuelong Wang and Jie Yang

Department of Mathematics, Computer Science, and Statistics
University of Illinois at Chicago

September 09, 2019

On the Agenda

Motivation of dimension reduction

Issues of high dimensional data (p is large)

- Curse of dimensionality (e.g. data points become sparse)
- Model overfitting

Two approaches

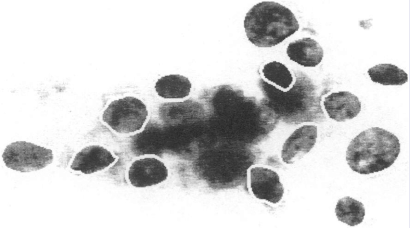
1. Variable selection
   - Forward/Backward selection, Shrinkage method (Lasso), etc.
2. **Dimension reduction** (Variable Projection)
   - Principle component analysis
   - Sufficient dimension reduction

An example: Breast cancer data

Data

- X: 30 Dependent variables are computed from a digitized image of a breast mass
- Y: Diagnosis results ($1 =$ malignant, $0 =$ benign)

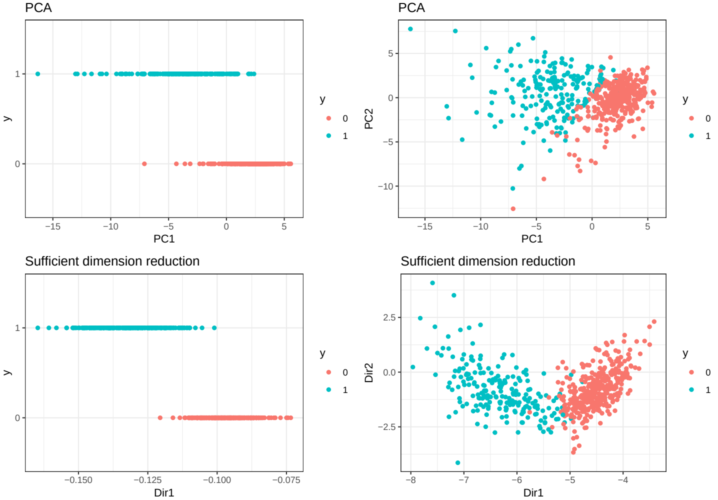
A Sample picture



Goal

Classification: Diagnose breast cancer from image-processed variables

An example: Breast cancer data



On the Agenda

Span and basis

Given d independent vectors $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$, $\mathbf{b}_i \in \mathbb{R}^p$,

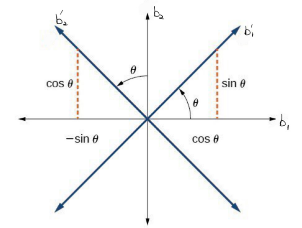
$$\text{Subspace } V = \mathcal{L}(\mathbf{b}_1, \ldots, \mathbf{b}_d) = \{\sum_{i=1}^{k} \lambda_i \mathbf{b}_i, \lambda_i \in \mathbb{R}\}$$

- $V = span(\mathbf{b}_1, \ldots, \mathbf{b}_d)$, V is spanned by B,
- $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$ is a basis of V

Basis is not unique

$$Span\begin{array}{cc} b_1 & b_2 \\ \end{array}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \Leftrightarrow Span\begin{array}{cc} b_1' & b_2' \\ \end{array}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Example



Sufficient dimension reduction

Fundamental assumption

Let random vector $X \in \mathbb{R}^{p\times 1}$, $Y \in \mathbb{R}$, $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d) \in \mathbb{R}^{p\times d}$, where $d << p$ and $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix.

$$Y|X \stackrel{d}{=} Y|B^T X$$

$$Y \perp\!\!\!\perp X|B^T X \Rightarrow Y \perp\!\!\!\perp X|(BA)^T X,$$

So B is not identifiable, but $span(B)$ is identifiable.

Sufficient dimension reduction

Dimension-reduction subspace (DRS)

$$Y \perp\!\!\!\perp X | P_S X, \quad P_S = B(B^T B)^{-1} B^T$$

$\mathcal{S}$ is called the dimension-reduction subspace.
However,$\mathcal{S}$ is not unique. Actually if $\mathcal{S} \subset \mathcal{S}_1$, then $\mathcal{S}_1$ is also a dimension-reduction space.

Target: Central Subspace

$$S_{Y|X} = \cap S_{DRS}$$

Under mild conditions, $S_{Y|X}$ is unique and a DRS subspace itself (Cook, 1996).

Take home message

- No model assumption between X and Y
- Target is a basis of the central subspace not specific values of coefficients
- A basis of subspace is $B = (\mathbf{b}_1, \ldots, \mathbf{b}_d)$

Estimating the central subspace

Principle component analysis (PCA)

1. $M = Var(X)$
2. Find the eigenvalues of M and arrange them in descending order $\lambda_1 \geq \ldots, \lambda_p$ and their corresponding eigenvectors $(u_1, \ldots, u_p)$
3. Select first several eigenvectors based on the total variation
4. $(\hat{u}_1, \ldots, \hat{u}_d) = (\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_d)$

Estimating the central subspace (cont.)

Sliced Inverse Regression (SIR) (Li 1991)

1. $Z = \Sigma_X^{-1/2}(X - E(X))$
2. $M_{SIR} := \Sigma_X^{1/2} Var(E(Z|Y))$
3. Find the eigenvalues and eigenvectors of M_{SIR}

Sliced Average Variance Estimation (SAVE) (Cook et al. 1991)

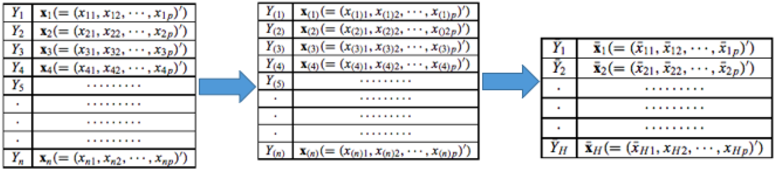
1. $Z = \Sigma_X^{-1/2}(X - E(X))$
2. $Var(Z|Y)$ is the conditional variance of X given Y
3. $M_{SAVE} := f(Var(Z|Y))$
4. Find the eigenvalues and eigenvectors of M_{SAVE}

How to estimate the $E(Z|Y)$, $Var(Z|Y)$?

1. Sort the data based on the response

$$Y_1 \dots, Y_n \Rightarrow Y^{(1)}, \dots, Y^{(n)}$$

2. Split data into H slices based on sorted $Y^{(i)}$
3. Within the slice h, calculate the $\hat{E}(Z|Y)$, $\hat{Var}(Z|Y)$,



Y_1	$\mathbf{x}_1(=(x_{11}, x_{12}, \cdots, x_{1p})')$
Y_2	$\mathbf{x}_2(=(x_{21}, x_{22}, \cdots, x_{2p})')$
Y_3	$\mathbf{x}_3(=(x_{31}, x_{32}, \cdots, x_{3p})')$
Y_4	$\mathbf{x}_4(=(x_{41}, x_{42}, \cdots, x_{4p})')$
Y_5	....
.	....
.	....
.	....
Y_n	$\mathbf{x}_n(=(x_{n1}, x_{n2}, \cdots, x_{np})')$

Original data

$Y_{(1)}$	$\mathbf{x}_{(1)}(=(x_{(1)1}, x_{(1)2}, \cdots, x_{(1)p})')$
$Y_{(2)}$	$\mathbf{x}_{(2)}(=(x_{(2)1}, x_{(2)2}, \cdots, x_{(2)p})')$
$Y_{(3)}$	$\mathbf{x}_{(3)}(=(x_{(3)1}, x_{(3)2}, \cdots, x_{(3)p})')$
$Y_{(4)}$	$\mathbf{x}_{(4)}(=(x_{(4)1}, x_{(4)2}, \cdots, x_{(4)p})')$
$Y_{(5)}$	....
.	....
.	....
.	....
$Y_{(n)}$	$\mathbf{x}_{(n)}(=(x_{(n)1}, x_{(n)2}, \cdots, x_{(n)p})')$

Sorted and sliced by y

$\bar{Y}_1$	$\bar{\mathbf{x}}_1(=(\bar{x}_{11}, \bar{x}_{12}, \cdots, \bar{x}_{1p})')$
$\bar{Y}_2$	$\bar{\mathbf{x}}_2(=(\bar{x}_{21}, \bar{x}_{22}, \cdots, \bar{x}_{2p})')$
.	....
.	....
.	....
$\bar{Y}_H$	$\bar{\mathbf{x}}_H(=(\bar{x}_{H1}, x_{H2}, \cdots, x_{Hp})')$

Slice means of standardized data

Issue with Binary response

- A binary response only has two levels, e.g. $0, 1$.
- Only two slices are available after slicing
- SIR can only find one direction

On the Agenda

Using conditional variance (Cook. 1999)

Main Idea

$\Delta = \Sigma_{X|Y=1} - \Sigma_{X|Y=0}$ could contain all the information of the central space

Not full rank

There are cases that $\hat{\Delta}$ is not full rank or even is 0 matrix

Probability Enhanced (PRE) method (Shin et al. 2014)

Main idea

- $S_{Y|X} = S_{G(X)|X}$, $G(x) = \mathcal{P}(Y = 1|X = x)$ is the conditional probability
- $Y \Rightarrow G(X) \in [0, 1]$
- Weighted Support Vector Machine(WSVM) to estimate the $\hat{G}(X)$

Computational time

- SVM method is sensitive to the number of observation N
- Tunning parameters

On the Agenda

Representative approach

Representative

A Representative is a summary statistic of data points within a cluster: For $(X_i, Y_i), i \in I_k$ and n_k is sample size of I_k

$$\bar{X}_k = R(X_1, \ldots, X_{n_k}) = \frac{\sum_i X_i}{n_k}, \quad \bar{Y}_k = R(Y_1, \ldots, Y_{n_k}) = \frac{\sum_i Y_i}{n_k},$$

where R is the summarizing function.

Steps

1. Cluster $(X_1, \ldots, X_N)$ into k groups $I_1, \ldots, I_k$, e.g.k-means
2. Calculate the representatives for each cluster I_k
3. Apply dimension reduction methods on the k representatives

How it works

Main idea

Y and $G(X)$ have identical central space: $S_{Y|X} = S_{G(X)|X}$

$$Y = f(\mathbf{b}_1^T X, \ldots, \mathbf{b}_d^T X, \epsilon) \Rightarrow \mathcal{P}(Y = 1|X) = G(\mathbf{b}_1^T X, \ldots, \mathbf{b}_d^T X)$$

For the Representative

$$\bar{Y}_k = \hat{\mathcal{P}}(Y = 1|X_i, i \in I_k) \approx G(\bar{X}_k) = G(\mathbf{b}_1^T \bar{X}_k, \ldots, \mathbf{b}_d^T \bar{X}_k)$$

Aysmptotic property

Let K be the total number of clusters, n_k be the total observations within cluster k, v_k be the cluster's volume.

Cluster with fixed volume

In this case, K and v_k are fixed, $n_k \to \infty$ as $N \to \infty$

$$\bar{Y}_k - G(\bar{\mathbf{X}}_k) \xrightarrow{P} \mu_g - G(\boldsymbol{\mu}_k) \neq 0$$

Cluster with shrinking volume

In this case, $K \to \infty$, $v_k \to 0, n_k \to \infty$ as $N \to \infty$

$$E([\bar{Y}_k - G(\bar{\mathbf{X}}_k)]^2) = O(N^{-\delta(r)})$$

- $K = O(N^{\frac{p}{4+p}})$

Additional value: Big data solution (N is large)

Clustering step

Clustering step reduced the sample size from N to K.

- $(Y_1, X_1) \ldots (Y_N, X_N) \to (\bar{Y}_1, \bar{X}_1) \ldots (\bar{Y}_k, \bar{X}_K)$

Parallel Algorithm for SIR and SAVE

1. Split the sliced data into b blocks, $X_1, \ldots X_B$
2. Load each block X_b and calculate the statistics for each block such as $\bar{X}_b, X_b^T X_b$
3. Summary the statistics across the blocks to get the candidate matrix M_{SIR}, M_{SAVE}

On the Agenda

Simulation setup

Data generation model: logit model

$$\log\left(\frac{\mathcal{P}(Y=1|X)}{\mathcal{P}(Y=0|X)}\right) = (\mathbf{b}_1^T X)^2 \cdot sin(\mathbf{b}_2^T X) \cdot exp(\mathbf{b}_3^T X)$$

- $X \in \mathbb{R}^6$
- $\mathbf{b}_i = \mathbf{e}_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^6$
- $S_{Y|X} = Span(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$
- $n = \{10^3, 10^4, 10^5, 10^6\}$

How to evaluate estimated central subspace

The number of direction

- Hypothesis Test: test if a eigenvalue is significant different than 0

Frobenius Distance

$$F = \|P_B - P_A\|_F$$

- $P_A = A(A^TA)^{-1}A$
- $\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$
- small value is better
- 0 means $Span(A) = Span(B)$

Trace correlation (R)

$$R = 1 - \frac{1}{k}\sum_{i=1}^{k} \rho_i^2$$

- ρ_i^2 is the eigenvalues of B^TAA^TB
- small value is better
- 0 means $Span(A) \subseteq Span(B)$

Result table

Table 1: Simulation result of table

		Method A	Method B
		log n	
	H_0 vs H_1		
Power	0D vs $>=$ 1D		
	1D vs $>=$ 2D		
	2D vs $>=$ 3D		
Type-I error	3D vs $>=$ 4D		
	4D vs $>=$ 5D		
	5D vs $>=$ 6D		
Distance	Frobenius		
	Trace		

Simulation result of SAVE

Table 2: Simulation result of SAVE
Significant level 0.05
directions of central subsapce $d = 3$

		Original SAVE				Proposed SAVE			
		log n							
	H_0 vs H_1	3	4	5	6	3	4	5	6
Power	0D vs $>=$ 1D	0.9	1	1	1	0	0.05	**1**	**1**
	1D vs $>=$ 2D	0.08	0.52	0.52	0.5	0	0	**1**	**1**
	2D vs $>=$ 3D	0	0.05	0.06	0.06	0	0	0.05	**1**

Simulation result of SAVE

Table 3: Simulation result of SAVE
Significant level 0.05
directions of central subsapce $d = 3$

		Original SAVE				Proposed SAVE			
		log n							
	H_0 vs H_1	3	4	5	6	3	4	5	6
Power	0D vs >= 1D	0.9	1	1	1	0	0.05	**1**	**1**
	1D vs >= 2D	0.08	0.52	0.52	0.5	0	0	**1**	**1**
	2D vs >= 3D	0	0.05	0.06	0.06	0	0	0.05	**1**
Type-I error	3D vs >= 4D	0	0	0	0.01	0	0	0	0.14
	4D vs >= 5D	0	0	0	0	0	0	0	0.03
	5D vs >= 6D	0	0	0	0	0	0	0	0.02

Simulation result of SAVE

Table 4: Simulation result of SAVE
Significant level 0.05
directions of central subsapce $d = 3$

		Original SAVE				Proposed SAVE			
		log n							
	H_0 vs H_1	3	4	5	6	3	4	5	6
Power	0D vs >= 1D	0.9	1	1	1	0	0.05	**1**	**1**
	1D vs >= 2D	0.08	0.52	0.52	0.5	0	0	**1**	**1**
	2D vs >= 3D	0	0.05	0.06	0.06	0	0	0.05	**1**
Type-I error	3D vs >= 4D	0	0	0	0.01	0	0	0	0.14
	4D vs >= 5D	0	0	0	0	0	0	0	0.03
	5D vs >= 6D	0	0	0	0	0	0	0	0.02
Distance	Frobenius	1.47	1.2	1.21	1.21	NA	1.44	**1.00**	**0.39**
	Trace	0.06	0.01	0.01	0.01	NA	0.02	**0.01**	**0.04**

Simulation result of SIR

Table 5: Simulation result of SIR

$$(\mathbf{b}_1^T x)^2 \cdot sin(\mathbf{b}_2^T x) \cdot exp(\mathbf{b}_3^T x)$$

		Original SIR				Proposed SIR			
		log n							
Power	Direction/Distance	3	4	5	6	3	4	5	6
	0D vs >= 1D	1	1	1	1	0.75	**1**	**1**	**1**
	1D vs >= 2D	NA	NA	NA	NA	0.16	**1**	**1**	**1**
	2D vs >= 3D	NA	NA	NA	NA	0.01	0.01	0	0.01

Simulation result of SIR

Table 6: Simulation result of SIR
Significant level 0.05
directions of central subsapce $d = 3$

		Original SIR				Proposed SIR			
		log n							
	Direction/Distance	3	4	5	6	3	4	5	6
Power	0D vs >= 1D	1	1	1	1	0.75	**1**	**1**	**1**
	1D vs >= 2D	NA	NA	NA	NA	0.16	**1**	**1**	**1**
	2D vs >= 3D	NA	NA	NA	NA	0.01	0.01	0	0.01
Type-I error	3D vs >= 4D	NA	NA	NA	NA	0	0	0	0
	4D vs >= 5D	NA	NA	NA	NA	0	0	0	0
	5D vs >= 6D	NA	NA	NA	NA	0	0	0	0

Simulation result of SIR

Table 7: Simulation result of SIR
Significant level 0.05
directions of central subsapce $d = 3$

		Original SIR				Proposed SIR			
		log n							
Power	Direction/Distance	3	4	5	6	3	4	5	6
	0D vs >= 1D	1	1	1	1	0.75	**1**	**1**	**1**
	1D vs >= 2D	NA	NA	NA	NA	0.16	**1**	**1**	**1**
	2D vs >= 3D	NA	NA	NA	NA	0.01	0.01	0	0.01
Type-I error	3D vs >= 4D	NA	NA	NA	NA	0	0	0	0
	4D vs >= 5D	NA	NA	NA	NA	0	0	0	0
	5D vs >= 6D	NA	NA	NA	NA	0	0	0	0
Distance	Frobenius	1.14	1.12	1.14	1.13	1.47	1.13	**1.01**	**1**
	Trace	0.01	0	0	0	0.06	0.02	**0**	**0**

On the Agenda

Conclusion and Future work

Pros

- Better recover the $S_{Y|X}$ in binary responses
  - Proposed SAVE can find all the basis of central space
  - Proposed SIR can find more then 1 direction as long as the directions are not symmteric
- Greatly shorten the running time in big data

Cons

- Need large sample ($N = 10^5$) to have accurate estimation
- Need to find a better hypothesis test for representative approach

Future work

- Apply our proposed method to a real dataset
- Combine the SDR method with classification methods

Reference

Cook, R Dennis, and Sanford Weisberg. 1991. "Discussion of 'Sliced Inverse Regression for Dimension Reduction'."

Kim, Boyoung, and Seung Jun Shin. 2019. "Principal Weighted Logistic Regression for Sufficient Dimension Reduction in Binary Classification."

Li, Ker-Chau. 1991. "Sliced Inverse Regression for Dimension Reduction."

Shin, Seung Jun, Yichao Wu, Hao Helen Zhang, and Yufeng Liu. 2014. "Probability-Enhanced Sufficient Dimension Reduction for Binary Classification."

Thank You

Backup

Examples

1. Linear regression: $Y = a + b_1^T X + b_2^T X + \epsilon$
2. NonLinear regression: $Y = a + \exp(b_1^T X) + \sin(b_2^T X) + \epsilon$
3. More general: $Y = f(b_1^T X, b_2^T X, \epsilon)$

Subspace

- Vector space U: $\vec{\mathbf{a}}, \vec{\mathbf{b}} \in U$
  1. $\vec{\mathbf{a}} + \vec{\mathbf{b}} \in U$
  2. $\lambda\vec{\mathbf{a}} \in U, \lambda \in \mathbb{R}$
- Subspace V: Given k independent vectors $(\vec{\mathbf{a}}_1, \ldots, \vec{\mathbf{a}}_k)$, $\vec{\mathbf{a}}_i \in \mathbb{R}^p$,

$$V = \mathcal{L}((\vec{\mathbf{a}}_1, \ldots, \vec{\mathbf{a}}_k) = \{\sum_{i=1}^{k} \lambda_i a_i, \lambda_i \in \mathbb{R}\}$$

  V is spaced by $(\vec{\mathbf{a}}_1, \ldots, \vec{\mathbf{a}}_k)$
- A basis of V: $(\vec{\mathbf{a}}_1, \ldots, \vec{\mathbf{a}}_k)$ is called a basis of V, but it is not unique

SIR

1. $E(X|Y) - E(X)$ is p-dimensional curves as Y varies and lies in a k-dimensional subspace
2. The covariance matrix of $E(X|Y) - E(X)$ is degenerate at any direction that orthogonal to $\Sigma_X b_i, i = 1, \ldots, d$
3. Candidate Matrix: $M_{SIR} = Var(E(X|Y) - E(X)) = Var(E(X|Y))$
4. $S_{SIR} := Span(\Sigma_X^{-1} M_{SIR}) \subseteq S_{Y|X}$
5. $\Sigma_X^{-1} M_{SIR} b_i = \lambda_i b_i$ b_i is the ith eigenvector of $\Sigma_X^{-1} M_{SIR}$

Simulation estimated direction: Proposed SAVE

	[,1]	[,2]	[,3]
[1,]	1.00	0.05	-0.02
[2,]	-0.02	-0.01	-1.00
[3,]	0.05	-1.00	0.01
[4,]	0.01	0.00	-0.01
[5,]	-0.02	0.00	-0.05
[6,]	0.00	0.01	0.02

Simulation estimated direction: Proposed SIR

```
      [,1]  [,2]
[1,]  0.00 -0.01
[2,] -0.01 -1.00
[3,] -1.00  0.01
[4,]  0.00 -0.03
[5,]  0.01 -0.01
[6,]  0.00  0.03
```

Simulation estimated direction: PRE SIR

```
      [,1]  [,2]  [,3]
[1,] -0.06 -0.01 -0.85
[2,]  0.01  0.98 -0.14
[3,] -0.97  0.00 -0.13
[4,]  0.03  0.26  0.11
[5,]  0.14 -0.13 -0.28
[6,]  0.20 -0.11 -0.35
```

Real Data analysis

SUSY data

- $n = 5 \times 10^7$
- X are 18 features of a physics experiment of particles in high-energy
- Y is binary

How evalute the estimated directions

- Don't know the true central space so no distance measure
- Classification performance but depends on the classification model