

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 03, 2019

- 1 Background
- 2 Existing solution
- 3 Our approach
- 4 Simulation Study
- 5 Conclusion

On the Agenda

1 Background

- Motivation
- SDR
- Estimating the central subspace

2 Existing solution

- PRE

3 Our approach

- Rep

4 Simulation Study

5 Conclusion

Motivation of reducing the dimension of the data

Curse of dimensionality

Two approaches

- 1 Variable selection (feature selection)
 - Lasso
- 2 **Dimension reduction** (feature projection)
 - Principle component analysis
 - Sufficient dimension reduction

An example: Breast Cancer data

Data

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass

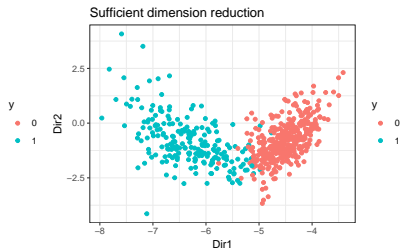
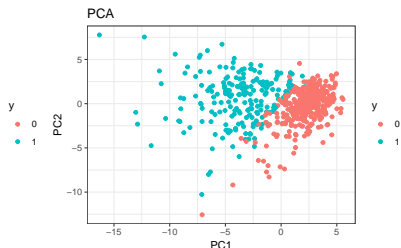
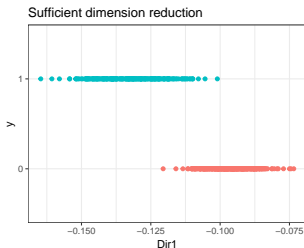
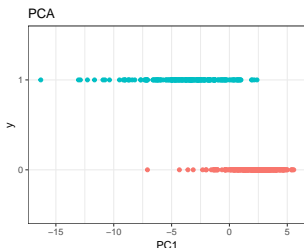
- Y: Diagnosis results (1 = malignant, 0 = benign)
- X: 30 real-valued features of each cell nucleus
 - e.g. radius, texture, area

picture

Goal

Classification: Diagnose breast cancer from image-processed nuclear features of fine needle aspirates

An example: Breast Cancer data



Subspace

- Vector space U : $\vec{a}, \vec{b} \in U$
 - 1 $\vec{a} + \vec{b} \in U$
 - 2 $\lambda \vec{a} \in U, \lambda \in \mathbb{R}$
- Span: Given $\mathbf{A} = (\vec{a}_1, \dots, \vec{a}_k)$, $\vec{a}_k \in \mathbb{R}^n$, a k -dimension subspace V spanned by \mathbf{A} is all the linear combinations of $(\vec{a}_1, \dots, \vec{a}_k)$

$$V = \mathcal{L}((\vec{a}_1, \dots, \vec{a}_k)) = \left\{ \sum_{i=1}^k \lambda_i \vec{a}_i, \lambda_i \in \mathbb{R} \right\}$$

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

Sufficient dimension reduction

Fundamental assumption

Let random vector $X \in \mathbb{R}^{p \times 1}$, $Y \in \mathbb{R}$, $B = (b_1, \dots, b_d) \in \mathbb{R}^{p \times d}$, where $d \ll p$ 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 $\text{span}(B)$ 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).

Estimating the central subspace

Sliced Inverse Regression (SIR) (Li 1991)

$$E(X|Y) - E(X) \in \Sigma_X S_{Y|X} = \text{Span}(b_k \Sigma_X), k = 1, \dots, K$$

- 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 $b_k \Sigma_X, k = 1, \dots, K$
- ③ Condinate Matrix: $M_{SIR} = \text{Var}(E(X|Y))$
- ④ $S_{SIR} := \text{Span}(\Sigma_X^{-1} M) \subseteq S_{Y|X}$
- ⑤ $\Sigma_X^{-1} M b = \lambda b \Rightarrow M b = \lambda \Sigma_X b$

Estimating the central subspace cont.

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

$$\text{span}(\Sigma_x - \Sigma_{X|\tilde{Y}}) \subseteq S_{Y|X} \Rightarrow \hat{B} = (\hat{b}_1, \dots, \hat{b}_d)$$

- There are many other methods using first and second moments together
 - Directional regression etc.

How to estimate the $E(X|Y)$?

- 1 Sort the data based on the response $Y_1 \dots, Y_n$
- 2 Split data into H slices based on the sorted responses $Y^{(1)}, \dots, Y^{(n)}$
- 3 Within each slice, replace the responses with slice label $h \in 1, \dots, H$
- 4 Within the slice h, calculate the average of X, $\tilde{X}_h = \hat{E}(X|Y = \tilde{Y}_h)$

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

Original data

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

Sorted and sliced by y

\tilde{Y}_1	$\tilde{\mathbf{x}}_1 = (\tilde{x}_{11}, \tilde{x}_{12}, \dots, \tilde{x}_{1p})'$
\tilde{Y}_2	$\tilde{\mathbf{x}}_2 = (\tilde{x}_{21}, \tilde{x}_{22}, \dots, \tilde{x}_{2p})'$
•
•
•
\tilde{Y}_H	$\tilde{\mathbf{x}}_H = (\tilde{x}_{H1}, \tilde{x}_{H2}, \dots, \tilde{x}_{Hp})'$

Slice means of standardized data

Issue with Binary response

Binary response only has two levels, e.g. 0, 1. So that only two slices are available

2 slices affects the performances of SDR method

- For SIR, it can only find one direction at most
- For SAVE, it also suffers from the limit number of slices

On the Agenda

1 Background

- Motivation
- SDR
- Estimating the central subspace

2 Existing solution

- PRE

3 Our approach

- Rep

4 Simulation Study

5 Conclusion

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

Main idea

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

Computational time

- SVM method is sensitive to the number of observation N

On the Agenda

1 Background

- Motivation
- SDR
- Estimating the central subspace

2 Existing solution

- PRE

3 Our approach

- Rep

4 Simulation Study

5 Conclusion

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

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

where R is the summarizing function.

Steps

- 1 Cluster (X_1, \dots, X_N) into K groups I_1, \dots, 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 $P(X)$ have identical central space: $S_{Y|X} = S_{P(X)|X}$

$$Y = f(b_1^T X, \dots, b_d^T X, \epsilon) \Rightarrow \\ \mathcal{P}(Y = 1|X) = P(X) = G(b_1^T X, \dots, b_d^T X)$$

For the Representative

$$\bar{Y}_k = \hat{\mathcal{P}}(Y = 1|X_i, i \in I_k) \approx G(b_1^T \bar{X}_k, \dots, b_d^T \bar{X}_k)$$

Aysmptotic property with fixed clusters

Fixed cluster

$$\begin{aligned}\bar{Y}_k - G(\bar{\mathbf{X}}_k) &\xrightarrow{P} \mu_g - G(\mu_k) \\ &= p_k^{-1} \int_{B_k} G(\mathbf{x}) F(d\mathbf{x}) - G\left(p_k^{-1} \int_{B_k} \mathbf{x} F(d\mathbf{x})\right)\end{aligned}$$

- Note that with fixed cluster, there is a bias between the representative version of conditional probability
- To remove the bias we need to reduce the size of cluster when N is increaseing

Asymptotic property with shrinking clusters

Shrinking cluster

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

Where $\delta(r) = \min\{4/(rd), 1 - 1/r\}$ for $r > 1$, which is maximized at $r = 1 + 4/d$. In other words, the minimum decreasing rate of $E([\bar{Y}_k - G(\bar{\mathbf{X}}_k)]^2)$ is $O(N^{-4/(d+4)})$ which is attained at $r = 1 + 4/d$.

Additional value: Big data solution (N is large)

Clustering step

Clustering step reduced the sample size from N to K .

- $(Y_1, X_1) \dots (Y_N, X_N) \rightarrow (Y_1^*, X_1^*) \dots (\bar{Y}_k, \bar{X}_k)$
- Note if the data set is too large, we could also use the online clustering method.

Additional value: Big data solution (N is large)

Parallel Algorithm for SIR and SAVE

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

On the Agenda

1 Background

- Motivation
- SDR
- Estimating the central subspace

2 Existing solution

- PRE

3 Our approach

- Rep

4 Simulation Study

5 Conclusion

Simulation setup

Data generation model: logit model

$$\log \left(\frac{\mathcal{P}(Y = 1|X = x)}{1 - \mathcal{P}(Y = 1|X = x)} \right) = b_1^T x \cdot \sin(b_2^T x) \cdot \exp(b_3^T x)$$

- $n = \{10^3, 10^4, 10^5, 10^6\}$ - $X \in \mathbb{R}^6$ -

$b_1 = e_i = (0, \dots, 1, \dots, 0) \in \mathbb{R}^6$ - $S_{Y|X} = \text{Span}(e_1, e_2, e_3)$

Note that the central subspace is a 3-dimensional subspace in a 6-dimensional space

How to evaluate esimated central subspace

The number of direction

- 1 Hypothesis Test: test if a eigenvalue is significant than 0
- 2 Ad-hoc: select all the eigenvalues whcih are larger then a cutoff value

The distance of the true subspace

- 1 Fourbin distance
- 2 trace correlation

Simulation result of SAVE

Table 1: Simulation result of SAVE

		Original SAVE				Proposed SAVE			
		log n							
	H_0 vs H_1	3	4	5	6	3	4	5	6
Power	0D vs \geq 1D	0.9	1	1	1	0	0.05	1	1
	1D vs \geq 2D	0.08	0.52	0.52	0.5	0	0	1	1
	2D vs \geq 3D	0	0.05	0.06	0.06	0	0	0.05	1
Type-I	3D vs \geq 4D	0	0	0	0.01	0	0	0	0.14
	4D vs \geq 5D	0	0	0	0	0	0	0	0.03
	5D vs \geq 6D	0	0	0	0	0	0	0	0.02
Distance	F	1.47	1.2	1.21	1.21	.	1.44	1.00	0.39
	R	0.06	0.01	0.01	0.01	.	0.02	0.01	0.04

Simulation result of SIR

Table 2: Simulation result of SIR

		SIR_Binary				SIR_PRE				SIR_R			
		log n											
Power	Direction/Distance	3	4	5	6	3	4	5	6	3	4	5	6
	0D vs \geq 1D	1	1	1	1	1	.	.	.	0.75	1	1	1
	1D vs \geq 2D	1	.	.	.	0.16	1	1	1
	2D vs \geq 3D	1	.	.	.	0.01	0.01	0	0.01
Type-I	3D vs \geq 4D	0	.	.	.	0	0	0	0
	4D vs \geq 5D	0	.	.	.	0	0	0	0
	5D vs \geq 6D	0	.	.	.	0	0	0	0
Distance	F	1.14	1.12	1.14	1.13	0.88	.	.	.	1.47	1.13	1.01	1
	R	0.01	0	0	0	0.06	.	.	.	0.06	0.02	0	0

On the Agenda

1 Background

- Motivation
- SDR
- Estimating the central subspace

2 Existing solution

- PRE

3 Our approach

- Rep

4 Simulation Study

5 Conclusion

Conclusion and Future work

Conclusion

- Better recover the central space in binary responses
- Greatly shorten the running time in big data

Future work

- Investigate optimal the choice of k to achieve the best performance of SDR 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."

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: $Y = f(b_1^T X, b_2^T X, \epsilon)$

SUSY data

SUSY data cont.