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 05, 2019

1. Motivation
2. Background and Issue
3. Existing solution
4. Our approach
5. Simulation Study
6. Conclusion

On the Agenda

1. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. Conclusion

Motivation of reducing the dimension of the data

Curse of dimensionality (p is large)

- Data becomes sparse (need more data to get same level of accuracy)
- Model Overfitting

Two approaches

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

An example: Breast Cancer data

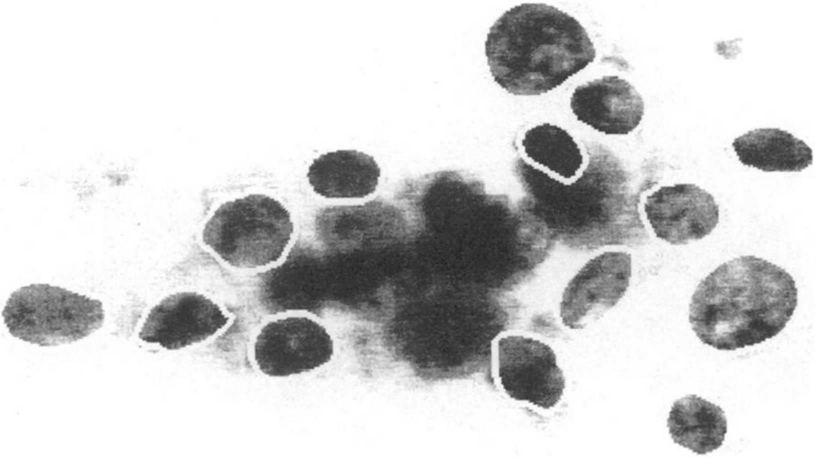


Figure 1:

An example: Breast Cancer data (Cont.)

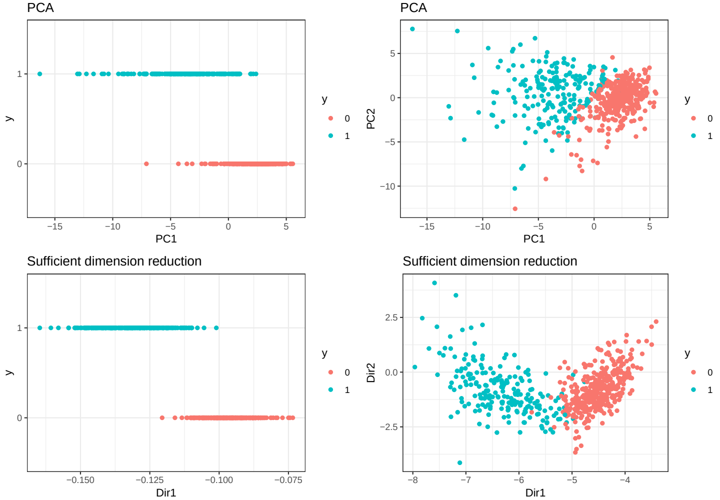
Data

- X: Dependent variables are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass
  - e.g. radius, texture, area
- Y: Diagnosis results (1 = malignant, 0 = benign)

Goal

Classificantion: Diagnose breast cancer from image-processed variables

An example: Breast Cancer data



PCA

PCA

Sufficient dimension reduction

Sufficient dimension reduction

On the Agenda

1. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. Conclusion

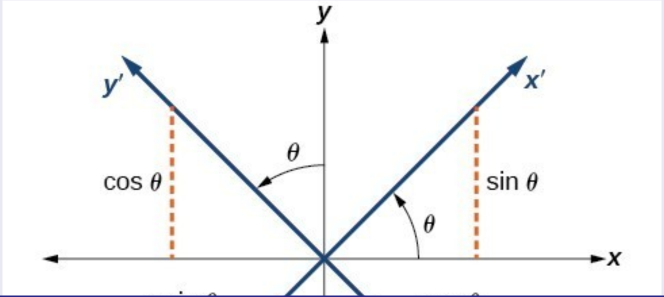
Span and basis

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

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

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

Basis is not unique



Sufficient dimension reduction

Fundamental assumption

Let random vector $X \in \mathbb{R}^{p\times 1}$, $Y \in \mathbb{R}$, $B = (b_1, \ldots, b_d) \in \mathbb{R}^{p\times d}$, where $d << p$ and $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix.

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

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

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

Sufficient dimension reduction

Dimension-reduction subspace (DRS)

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

$\mathcal{S}$ 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 vector)
- A basis of subspace is $B = (\vec{\mathbf{b}}_1, \ldots, \vec{\mathbf{b}}_d)$

Estimating the central subspace

Principle component analysis (PCA)

1. $M = \hat{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T$
2. Fine the eigenvalues of M and arrange them in decending 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

Note that

$$(u_1, \ldots, u_d) = (\hat{b}_1, \ldots, \hat{b}_d)$$

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)

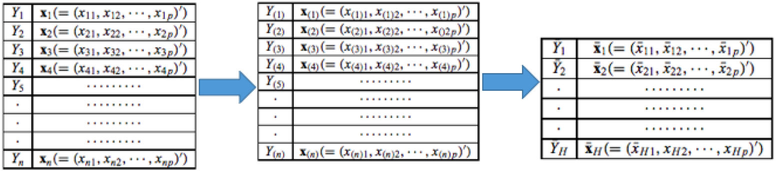
1. $Z = \Sigma_X^{-1/2}(X - E(X))$
2. $Var(Z|Y)$ is the conditional variance of X given Y
3. $M_{SAVE} := f(Var(Z|Y))$
4. Find the eigenvalues and eigenvectors of M_{SAVE}

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

1. Sort the data based on the response

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

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

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

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

- 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

1. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. Conclusion

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

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

Main idea

- $S_{Y|X} = S_{G(X)}$, $G(x) = \mathcal{P}(Y = 1|X = x)$ 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

1. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. 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

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

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: $S_{Y|X} = S_{G(X)|X}$

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

For the Representative

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

Aysmptotic property with fixed clusters

Fixed cluster

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

- 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

Aysmptotic property with shrinking clusters

Shrinking cluster

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

- $\delta(r)$ is maximized 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. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. Conclusion

Simulation setup

Data generation model: logit model

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

- $n = \{10^3, 10^4, 10^5, 10^6\}$
- $X \in \mathbb{R}^6$
- $b_1 = e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^6$
- $S_{Y|X} = Span(e_1, e_2, e_3)$

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

- Hypothesis Test: test if a eigenvalue is significant than 0
- Total Variance: $T = \frac{\sum_i^d \lambda_i}{\sum_j^p \lambda_j}$

Frobenius Distance (F)

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

where $P_A = A(A^T A)^{-1} A$

$\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$

right

1.
2. Trace correlation

$$r^2 = \frac{1}{k} \sum^k \rho_i^2$$

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 $>=$ 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	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	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 >= 1D	1	1	1	1	1	.	.	.	0.75	**1**	**1**	**1**
	1D vs >= 2D	.	.	.	.	1	.	.	.	0.16	**1**	**1**	**1**
	2D vs >= 3D	.	.	.	.	1	.	.	.	0.01	0.01	0	0.01
Type-I	3D vs >= 4D	.	.	.	.	0	.	.	.	0	0	0	0
	4D vs >= 5D	.	.	.	.	0	.	.	.	0	0	0	0
	5D vs >= 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. Motivation
   - Motivation
2. Background and Issue
   - SDR
   - Estimating the central subspace
3. Existing solution
   - Variance matrix
   - PRE
4. Our approach
   - Rep
5. Simulation Study
6. 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: $Y = f(b_1^T X, b_2^T X, \epsilon)$

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 $(\vec{\mathbf{a}}_1, \dots, \vec{\mathbf{a}}_k), \quad \vec{\mathbf{a}}_i \in \mathbb{R}^p,$

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

  V is spaced by $(\vec{\mathbf{a}}_1, \dots, \vec{\mathbf{a}}_k)$
- A basis of V: $(\vec{\mathbf{a}}_1, \dots, \vec{\mathbf{a}}_k)$ 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. Condidate 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 $\Sigma_X^{-1} M_{SIR}$

SUSY data

SUSY data cont.