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 04, 2019

On the Agenda

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 (feature selection)
   - Forward/Backward selection, Lasso, etc.
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

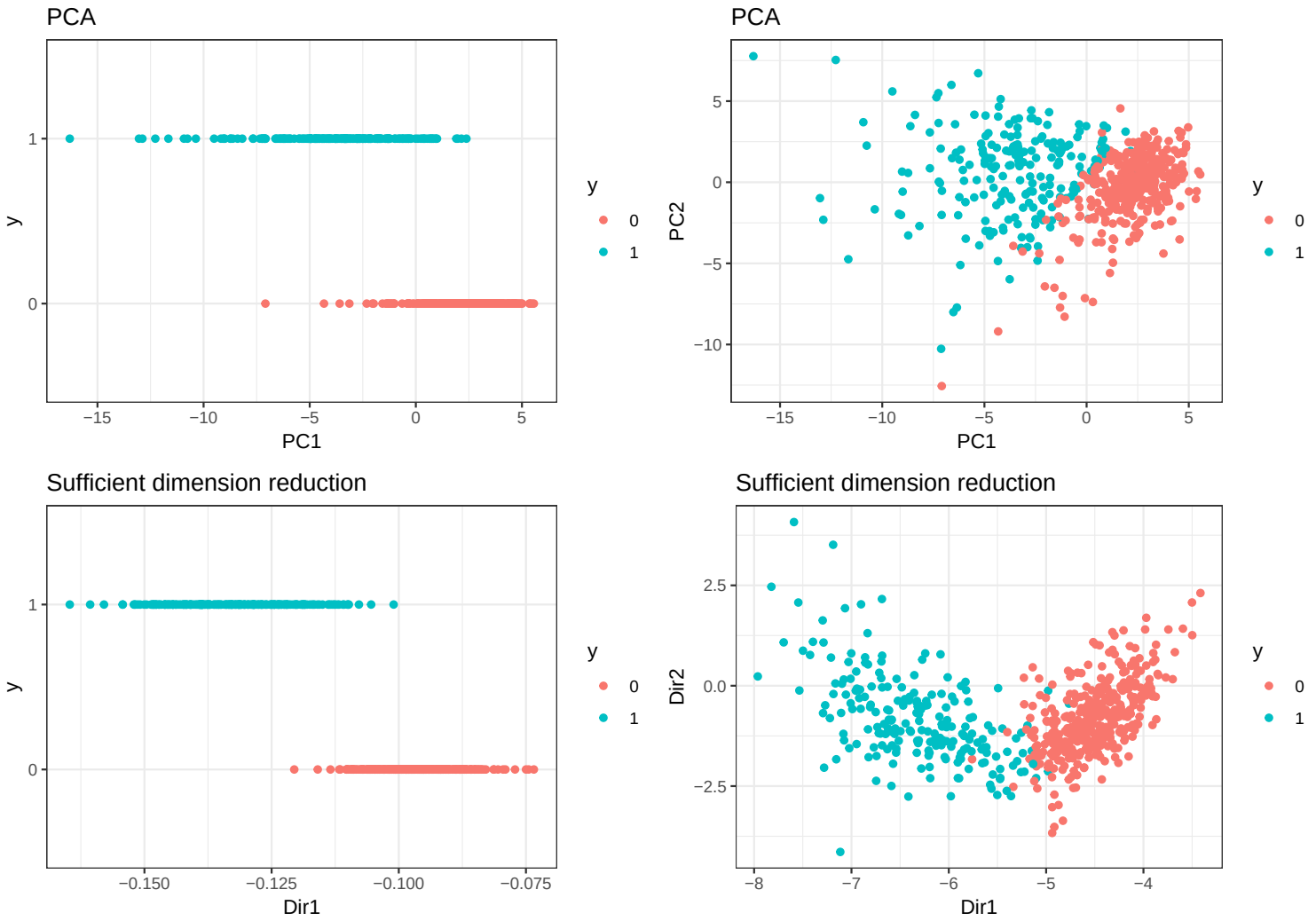
- Y: Diagnosis results (1 = malignant, 0 = benign)
- X: 30 features of each each cell nucleus
  - e.g. radius, texture, area

picture

Goal

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

An example: Breast Cancer data



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

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.

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

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

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

Sufficient dimension reduction

Dimension-reduction subspace (DRS)

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

$\mathcal{S}$ 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).

Comment

- No model assumption between X and Y
- Target is a subspace not a specific values coefficients

Estimating the central subspace

Sliced Inverse Regression (SIR) (Li 1991)

$$E(X|Y) - E(X) \in \Sigma_X S_{Y|X} = Span(\Sigma_X b_i), i = 1, \ldots, d$$

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 $\Sigma_X^{-1} M_{SIR}$

Estimating the central subspace (cont.)

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

$$span(\Sigma_x - \Sigma_{X|\tilde{Y}}) \subseteq S_{Y|X} \Rightarrow (b_1, \ldots, b_d)$$

- There are many other methods using first and second monments togehter
  - Directional regression etc.

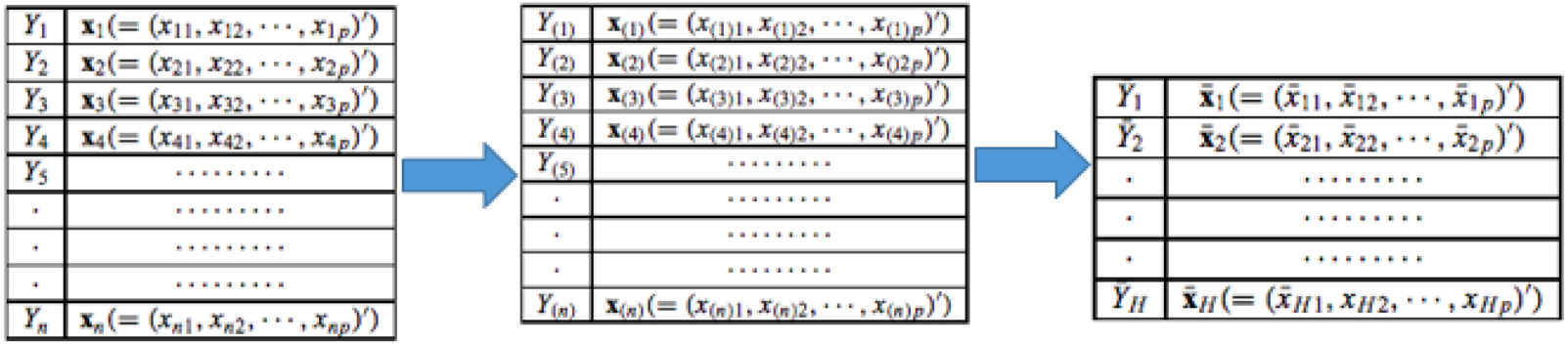
How to estimate the $E(X|Y)$, $\Sigma_{X|\tilde{Y}}$?

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 and set $Y = \tilde{Y}_h, h = 1, \ldots H$
3. Within the slice h, calculate the average of X,

$$\tilde{X}_h = \hat{E}(X|Y = \tilde{Y}_h)$$



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

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

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

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

- 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

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

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.

- $(Y_1, X_1) \ldots (Y_N, X_N) \rightarrow (Y_1^*, X_1^*) \ldots (\bar{Y}_k, \bar{X}_k)$
- 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, $X_1, \ldots X_B$
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

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

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 distrance 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 >= 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											
	Direction/Distance	3	4	5	6	3	4	5	6	3	4	5	6
Power	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

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.