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

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).

Estimating the central subspace

Inverse regression: Condition X on Y

To Estimate a linear subspace $\Rightarrow$ a Basis B of $S_{Y|X}$
Sliced Inverse Regression (SIR) (Li 1991)

$$E(X|Y) - E(X) \in \Sigma_X S_{Y|X} \Rightarrow \hat{B} = (\hat{b}_1, \ldots, \hat{b}_d)$$

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

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

Slicing method

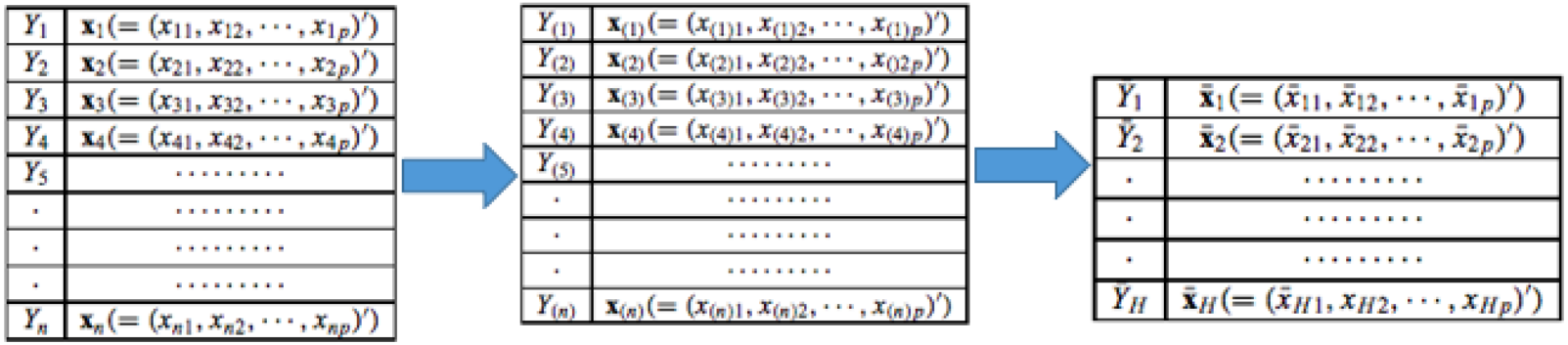

Figure 1:

1. Sort the data based on the response values.
2. Split data into the slices based on the sorted responses.

Binary response

Binary response only has two levels, e.g. $0, 1$.

Limited number of slices

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

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

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

$$X_k^* = R(X_1, \ldots, X_{n_k}) = \frac{\sum_i X_i}{n_k}, \quad Y_k^* = R(Y_1, \ldots, Y_{n_k}) = \frac{\sum_i Y_i}{n_k},$$

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

Additional value: Big data solution (N is large)

Clustering step

Clustering step reduced the sample size from N to K.

- $(Y_1, X_1) \ldots (Y_N, X_N) \rightarrow (Y_1^*, X_1^*) \ldots (Y_K^*, X_K^*)$
- 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}

Simulation setup

Data generation model: Latent model

$$Y = \begin{cases} 0 & f(b_1^T X, b_2^T X, b_3^T X, \epsilon) < 0 \\ 1 & \text{Otherwise} \end{cases}$$

where

- $X \in \mathbb{R}^6 \sim N(\mathbf{0}, \mathbf{I})$
- $b_i = e_i = (0, \ldots, 1, 0, \ldots, 0)^T$, so $b_1^T X = X_1, b_2^T X = X_2, b_3^T X = X_3$
- $\epsilon \sim N(0, 1)$

Simulation result

Performance evaluation

1. The number of directions of the central space: Hypothesis Test
2. Difference between a true bias B and an estimated $\hat{B}$:
   - Trace correlation and Frobenius distance

Result summary

- The true basis is (e_1, e_2, e_3).
- The proposed method is able to recover the whole true central space.
- Other methods can only find part of the central space.

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	0.01
	4D vs >= 5D	0	0	0	0	0	0	0	0
	5D vs >= 6D	0	0	0	0	0	0	0	0
Distance	F	1.33	1.2	1.21	1.19	1.71	1.03	**0.23**	**0.07**
	R	0.17	0.14	0.14	0.13	0.29	0.1	**0.01**	**0**

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

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

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	.	.	.	.	0.96	.	.	.	0.01	0.01	0	0.01
Type-I	3D vs >= 4D	.	.	.	.	0.5	.	.	.	0	0	0	0
	4D vs >= 5D	.	.	.	.	0.1	.	.	.	0	0	0	0
	5D vs >= 6D	.	.	.	.	0.01	.	.	.	0	0	0	0
Distance	F	1.13	1.05	1.06	1.09	1.14	.	.	.	1.37	1.29	1.24	1.29
	R	0.14	0.14	0.14	0.15	0.12	.	.	.	0.18	0.15	0.15	0.15

Iteration time is 200 and significant level is 0.05

Performance evaluation

1. The number of directions of the central space: Hypothesis Test
2. Difference between a true bias B and an estimated $\hat{B}$:

Simulation setup

Data generation model: Latent model

$$Y = \begin{cases} 0 & (b_1^T X)^2 * e^{(b_2^T X)} * \sin(b_3^T X) + \epsilon < 0 \\ 1 & \text{Otherwise} \end{cases}$$

where

- $X \in \mathbb{R}^6 \sim N(\mathbf{0}, \mathbf{I})$
- $b_i = e_i = (0, \ldots, 1, 0, \ldots, 0)^T$, so $b_1^T X = X_1, b_2^T X = X_2, b_3^T X = X_3$
- $\epsilon \sim N(0, 1)$
- $P(X) = \Phi((b_1^T X)^2 * e^{(b_2^T X)} * \sin(b_3^T X))$, where Φ is the CDF of standard normal distribution.

How it works

Main idea

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

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

For the Representative

$$Y_k^* = \hat{\mathcal{P}}(Y = 1|X_i, i \in I_k) \approx P(b_1^T \bar{X}_k, \ldots, b_d^T \bar{X}_k)$$
$$Y_k^* \to P(Y = 1|X = x_k) \text{ as } N, K, N/K \to \infty$$

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."