Representative approach for big data dimension reduction with binary responses

Xuelong Wang, Jie Yang

May 16, 2019

Sufficient Dimension Reduction

Model Settings

$$Y = f(X, \epsilon) = f(\beta_1 X, \ldots, \beta_k X, \epsilon)$$

x is explanatory variable, column vectors on $\mathcal{R}^p$,
$\beta' s$ are unknown row vectors,
ϵ is independent of X,
f is an arbitrary unknown function on $\mathcal{R}^{k+1}$

- $(\beta_1 X, \ldots, \beta_k X)'$ is the projection of the $X \in \mathcal{R}^p$ into $\mathcal{R}^K$, $K << p$
- Lower dimension projection of X contains most of the information

Sufficient Dimension Reduction

The space spaned by $\beta's$

1. Effective dimension-reduction direction (e.d.r)
   - A Linear combination of $\beta's$
2. A Linear space $\mathcal{B}$:
   - Spanned by $\beta's$ ($Span(\beta)$) $\Leftrightarrow$ All the possible linear combination of $\beta's$

- Since f is arbitrary, f and $\beta's$ are not Estimable
- Only the $\mathcal{B}$ can be identified
- Inverse Regression is one of the methods of estimating the Effective dimension-reduction space ($\mathcal{B}$)

Sliced Inverse Regression method

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

$$\hat{V} = n^{-1}\sum_{h=1}^{H} n_h \bar{x}_h \bar{x}_h^T$$

Estimated Covariance matrix

Conduct PCA on $\hat{V}$
Find the first Kth eigenvectors $\hat{\eta}$

$$\hat{\beta}_k = \hat{\eta}_k \Sigma_{xx}^{-1/2}$$

Binary response

The curse of Binary response for the sliced-based methods

Since the responses only have to values, the number slices

- For the method using only first moment, only one direction can be recovered
- For the method using more than moment(SAVE), there are situations they cannot recover all the directions

Existing apporaches

- Using SVM to estimate the sudo conditional probability

Representative approach

Data generating models for binary response

Let Y^* as the latent response and Y as the observed binary response,

$$Y = \begin{cases} 0 & Y^* < \theta \\ 1 & Y^* \geq \theta \end{cases}$$

Where θ is the cutoff value.

$$Y^* = f(X^T\beta_1, \ldots, X^T\beta_p, \epsilon),$$

Where ϵ is a random variable. Thus, we have

$$\mathcal{P}(Y = 1|X) = Pr(Y^* > \theta|X).$$

Note that different distribution ϵ will eventually affect the distribution of Y.

Data generating models

Conditional probability model

$$
\begin{aligned}
\lambda(\mathcal{P}(Y=1|X)) &= h(\beta_1^T X, \ldots, \beta_p X) \\
\mathcal{P}(Y=1|X) &= \lambda^{-1} \circ h(\beta_1^T X, \ldots, \beta_p X) \\
&= g(\beta_1^T X, \ldots, \beta_p X)
\end{aligned}
$$

Troubles for traditional methods

Simulation results

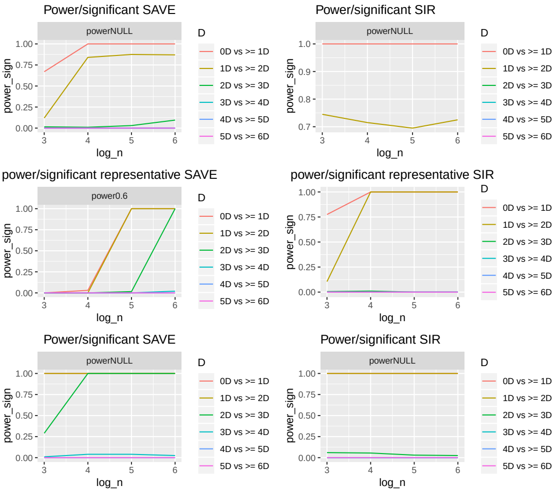

Figure 2:

Simulation results

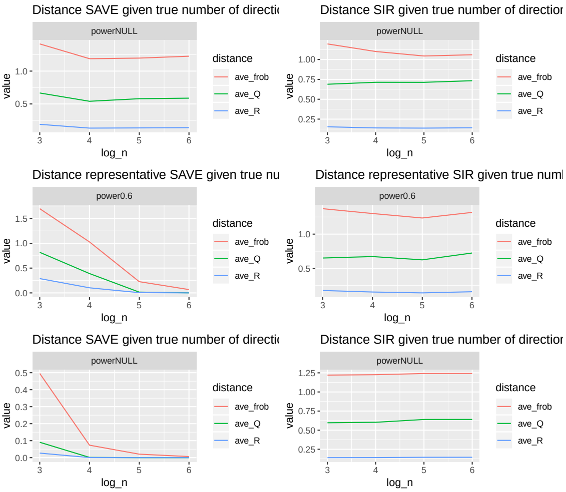

Figure 3:

The conditional expecation and representative

The choice of number of cluster K