Sliced Inverse Regression for Dimension Reduction

Ker-Chau Li

JASA, 1991

Outline

Introduction

- Why dimension reduction?
  - MLE for parametric model
  - Then nonparametric model for more flexible assumptions
  - Local smoothing in nonparametric theme
  - Dimensionality grows $\rightarrow$ need dimension reduction
- In this paper, consider the ideal case

$$y = f(\boldsymbol{\beta}_1^T \mathbf{x}, \boldsymbol{\beta}_2^T \mathbf{x}, \ldots, \boldsymbol{\beta}_K^T \mathbf{x}, \epsilon) \tag{1}$$

- $\boldsymbol{\beta}$'s unknown, ϵ independent of $\mathbf{x}$, f arbitrary unknown function

Introduction

- When K is small, we achieve the goal of data reduction
- Call the linear space $\mathcal{B}$ generated by the β's the effective dimension reduction (e.d.r.) space, and any direction within $\mathcal{B}$ the e.d.r. direction
- This serves as a pre-analysis step of data reduction
- Goal in this paper: estimate the e.d.r. directions
- Method: inverse regression, regress $\mathbf{x}$ against y
- $E(\mathbf{x}|y)$ coincides with the e.d.r. space $\mathcal{B}$, under some conditions

A model for dimension reduction

- This section describes what to implement after SIR data reduction
- Let $\sum_{\mathbf{xx}}$ be the covariance matrix of $\mathbf{x}$. Let $\mathbf{z} = \sum_{\mathbf{xx}}^{-1/2}(\mathbf{x} - E(\mathbf{x}))$ be the standardized version. Let $y = f(\boldsymbol{\eta}_1^T \mathbf{z}, \ldots, \boldsymbol{\eta}_K^T \mathbf{z}, \epsilon)$, where $\boldsymbol{\eta}_k = \sum_{\mathbf{xx}}^{1/2} \boldsymbol{\beta}_k$. Call the space generated by the $\boldsymbol{\eta}$'s a standardized e.d.r. space
- The criterion to evaluate the effectiveness of an estimated e.d.r. direction, which is invariant under scale change

$$R^2(\mathbf{b}) = \max_{\beta \in \mathcal{B}} \frac{(\mathbf{b}^T \sum_{\mathbf{xx}} \boldsymbol{\beta})^2}{\mathbf{b}^T \sum_{\mathbf{xx}} \mathbf{b} \cdot \boldsymbol{\beta}^T \sum_{\mathbf{xx}} \boldsymbol{\beta}} \tag{2}$$

The inverse regression curve

- **Condition 3.1** For any $\mathbf{b}$, the conditional expectation $E(\mathbf{b}^T\mathbf{x}|\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x})$ is linear in $\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x}$. That is, for some constants, $c_0, c_1, \ldots, c_K$,
$E(\mathbf{b}^T\mathbf{x}|\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x}) = c_0 + c_1\boldsymbol{\beta}_1^T\mathbf{x} + \cdots + c_K\boldsymbol{\beta}_K^T\mathbf{x}$
- The above condition is satisfied when the distribution of $\mathbf{x}$ is elliptically symmetric (normal distribution)
- **Theorem 3.1** Under (1) and Condition 3.1, the centered inverse regression curve $E(\mathbf{x}|y) - E(\mathbf{x})$ is contained in the linear subspace spanned by $\sum_{\mathbf{xx}}\boldsymbol{\beta}_k$
- Corollary about the standardized version

The inverse regression curve

- An important consequence: $\mathrm{cov}[E(\mathbf{z}|y)]$ is degenerate in any direction orthogonal to $\mathcal{B}$
- Thus, the eigenvectors associated with the largest K eigenvalues of $\mathrm{cov}[E(\mathbf{z}|y)]$ are the standardized e.d.r. directions
- By law of total variation, $E[\mathrm{cov}(\mathbf{z}|y)] = I - \mathrm{cov}[E(\mathbf{z}|y)]$, we can estimate $E[\mathrm{cov}(\mathbf{z}|y)]$ by the SIR algorithm introduced later, and eigenvalue decompose it. Then the eigenvectors associated with the smallest K eigenvalues are the e.d.r. directions

Sliced inverse regression

- The SIR algorithm, based on the sample $(y_i, \mathbf{x}_i); i = 1, \ldots, n$, is
  1. Standardize $\mathbf{x}$
  2. Divide range of y into H slices, $I_1, \ldots, I_H$; let the proportion of the y_i that falls in slice h be $\hat{p}_h$. That is, $\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_h(y_i)$, where $\delta_h(y_i)$ is the indicator function of whether y_i falls into the h^{th} slice or not.
  3. Within each slice, compute the sample mean of the $\mathbf{x}_i$'s, $\hat{\mathbf{m}}_h$.
  4. Conduct a (weighted) PCA for the data $\hat{\mathbf{m}}_h$ in the following way: Form the weighted covariance matrix $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{\mathbf{m}}_h \hat{\mathbf{m}}_h^T$, then find the eigenvalues and the eigenvectors for $\hat{V}$.
  5. Let the K largest eigenvectors be $\hat{\boldsymbol{\eta}}_k$. Output $\hat{\boldsymbol{\beta}}_k = \hat{\sum}_{\mathbf{xx}}^{-1/2} \hat{\boldsymbol{\eta}}_k$.

Sampling properties

- It is shown the output of SIR provides a root n consistent estimates for the e.d.r. directions, in this section
- The method used is basic statistics
- A simple approximation to $R^2(\mathbf{b})$ is given by

$$E[R^2(\hat{\mathcal{B}})] = 1 - \frac{p-K}{n}(-1 + \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\lambda_k}) + o(\frac{1}{n}) \tag{3}$$

- In practice, substitute the k^{th} largest eigenvalue of $\hat{V}$ for λ_k
- **Theorem:** If $\mathbf{x}$ is normally distributed, then $n(p-K)\bar{\lambda}_{(p-K)}$ asymptotically follows a χ^2 distribution with df $(p-K)(H-K-1)$. Here $\bar{\lambda}_{(p-K)}$ is the average of the smallest $p-K$ eigenvalues of V

Simulation results

- **Behavior of the SIR estimates**
- Model 1 setup: $K = 1$, $y = x_1 + x_2 + x_3 + x_4 + 0x_5 + \epsilon$
- $n = 100$, $p = 5$, ϵ follows $i.i.d.$ normal distribution, independent of $\mathbf{x}$
- 100 replicates, $H = 5, 10, 20$
- $\boldsymbol{\beta} = (0.5, 0.5, 0.5, 0.5, 0)^T$

Table 1. Mean and Standard Deviation of $\hat{\beta}_1 = (\hat{\beta}_{11}, \ldots, \hat{\beta}_{15})$ for the linear model (6.1), n = 100; the Target is (.5, .5, .5, .5, 0)*

H	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{13}$	$\hat{\beta}_{14}$	$\hat{\beta}_{15}$
5	.505	.498	.494	.488	.002
	(.052)	(.049)	(.056)	(.056)	(.066)
10	.502	.500	.492	.491	.001
	(.046)	(.045)	(.055)	(.049)	(.060)
20	.500	.502	.497	.487	−.003
	(.048)	(.046)	(.053)	(.054)	(.060)

Simulation results

- **Behavior of the SIR estimates**
- Model 2 setup: $y = x_1(x_1 + x_2 + 1) + \sigma \cdot \epsilon$
- $n = 400$, $p = 10$, the rest x_i's follow $i.i.d.$ normal distribution, independent of x_1 and x_2
- $\sigma = 0.5$ and 1
- Model 3 setup: $y = \frac{x_1}{0.5+(x_2+1.5)^2} + \sigma \cdot \epsilon$

Simulation results

Table 2. Mean and Standard Deviation of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ for the Quadratic Model (6.2), $p = 10$, $n = 400$

	$\sigma = 0.5$		$\sigma = 1$	
H	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.91	.75	.88	.52
	(.05)	(.15)	(.07)	(.21)
10	.92	.80	.89	.55
	(.04)	(.13)	(.08)	(.24)
20	.93	.77	.88	.49
	(.04)	(.15)	(.08)	(.26)

Figure:

Simulation results

- **Eigenvalues**
- Recall the only theorem in the paper
- The big table 4 in the paper assures that under the normal distribution, the theorem can be applied in practice to help determine what K should be, after the SIR algorithm.

Simulation results

- **Graphics**
- Due to lack of powerful plot tools, I cannot tell too much from the figures in the paper...