

# Lab Meeting

## Text Mining (SIR)

Choi Taeyoung

Sungkyunkwan University  
Department of Statistic

April 15, 2020



# Index

1 Introduce

2 Algorithm

3 Text Mining

4 Sliced inverse Regression

5 Future Study



# Introduce



# Sliced Inverse Regression(분할 역회귀)

## SIR의 개요

- SIR은  $R^p$ 공간 상에서 설명변수의 분포가 정규분포를 따르며 설명변수들 사이의 선형조건을 만족한다는 가정을 적용
- 설명변수의 선형결합인 SIR의 설명변수를 찾는 차원축소 방법
- 만약 설명변수  $X$ 의 선형결합 수가 알려져 있지 않다면, SIR는 설명변수에 대한 가정을 적용하여 반응변수  $y$ 를 효과적으로 찾을 수 있다고 알려져 있음



# 문제점

## 고려할 점

- SIR에서 유도되는 선형결합  $\beta^T X$ 가 원자료  $X$ 의 모든 설명변수에 관한 선형결합으로 표현되기 때문에
- $\eta^T = (\eta_1, \dots, \eta_d)$ 를 해석하고 적용시키는 데 있어서 의미있는 SIR의 방향 수를 적절히 결정하는데 문제가 발생
- 적절한 차원 축소 방법이 병행되어야한다. 예를 들면 LASSO( $L^1$  penalty)에 SIR을 적용시켜 해결
- $\Sigma_{xx}$ 는  $x$ 의 공분산행렬이고,  $z = \Sigma_{xx}^{-\frac{1}{2}} [x - E(x)]$ 는  $x$ 의 표준화라고 하자.
- 여기서  $\eta_k = \beta_k \Sigma_{xx}^{\frac{1}{2}} (k = 1, \dots, K)$ , 즉, 차원이 축소된 상태의  $X$ 와 같은 차원의 벡터



# Algorithm



# SIR

## 데이터와 노테이션

- $(y_i, x_i)$  ( $i = 1, \dots, n$ )인 데이터가 있다.
- $\bar{x}_i = \hat{\Sigma}_{xx}^{-1/2}(x_i - \bar{x})$  ( $i = 1, \dots, n$ )
- 이 때,  $\hat{\Sigma}_{xx}, \bar{x}$ 는 각각 표본 공분산행렬 그리고 표본평균이다.
- $y$ 의 범위로 분할된  $H$ 분할 영역인  $I_1, \dots, I_H$ , 여기서 분할  $s$ 로 떨어지는  $y_i$ 의 비율은  $\hat{p}_s$ 라고 한다.
- 여기서  $\hat{p}_s = (1/n) \sum_{i=1}^n \delta_h(y_i)$ , 또한 여기서  $\delta_s(y_i)$ 는  $y$ 가  $s$ -번째 분할  $I_s$ 에 속하는지 여부에 따라 0 또는 1 값을 취한다.



# SIR 알고리즘

- 1  $y$ 의 범위로 분할된  $\hat{p}_s$  비율로 구한 관측값 수를  $1/n\hat{p}_s$ 로 한 뒤 관측값들의 표본 분할평균  $\hat{m}_s = (1/n\hat{p}_s) \sum_{y_i \in I_s} x_i$ 을 구한다.
- 2 다음과 같은 가중 표본 분산-공분산 행렬  $\hat{V}$ 의 고유값분해를 통해 고유값  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_d$ 와 고유벡터  $\hat{\eta}_j, j = 1, 2, \dots, d$ 를 찾는다.  $\hat{V} = \sum_{h=1}^H \hat{p}_s \hat{m}_s \hat{m}_s'$
- 3  $K$  최대 고유벡터(행벡터)를  $\hat{\eta}_k$ 라고 설정하면,  $\hat{\beta}_k = \hat{\eta}_k \Sigma_{xx}^{-1/2} (k = 1, \dots, K)$
- 4 이때 행렬  $\hat{V}$ 의 고유값과 고유벡터는  $\sqrt{n}$ 의 비율로

$$Var[E(X|\tilde{Y})] = \sum_{s=1}^H Pr[\tilde{Y} = s] \mu_{X|s} \mu_{X|s}^T ; \mu_{X|s} = E[X|\tilde{Y} = s]$$

인  $[Var[E(X|\tilde{Y})]]$ 의 고유값  $\lambda_j$ 와 고유벡터  $\eta_j$ 에 수렴하게 된다.





# SIR 알고리즘

관측값의 수가  $n, p$ 차원 설명변수  $X$ 와 반응변수  $y$ 가 결합된  $(p+1)$ 차원의 원자료가 주어졌을 때 SIR의 알고리즘

[Step. 1] 반응변수  $y$ 의 크기를 기준으로 원자료를 순서화(sorting)한다

$$\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \left| \begin{array}{l} X_1 = (x_{11}, x_{12}, \dots, x_{1p}) \\ X_2 = (x_{21}, x_{22}, \dots, x_{2p}) \\ \vdots \\ X_n = (x_{n1}, x_{n2}, \dots, x_{np}) \end{array} \right.$$

[표 2-1] 원자료



# SIR 알고리즘

[Step. 2] 반응변수  $y$ 의 범위에 따라 원자료를  $H$ 개의  $I_1, I_2, \dots, I_H$ 로 분할한다

$$\begin{array}{l} I_1 \left\{ \begin{array}{l} y_{(1)} \\ y_{(2)} \\ \vdots \end{array} \right. \left| \begin{array}{l} X_{(1)} = (x_{(1)1}, x_{(1)2}, \dots, x_{(1)p}) \\ X_{(2)} = (x_{(2)1}, x_{(2)2}, \dots, x_{(2)p}) \\ \vdots \end{array} \right. \\ I_H \left\{ \begin{array}{l} y_{(n)} \end{array} \right. \left| \begin{array}{l} X_{(n)} = (x_{(n)1}, x_{(n)2}, \dots, x_{(n)p}) \end{array} \right. \end{array}$$

[표 2-2]  $y$ 에 의한 순서화 및  $H$  분할

[Step. 3] 각 분할  $s$ 에서의 분할평균(sliced mean)을 계산한다

$$\hat{m}_s = (1/n\hat{p}_s) \sum_{y_i \in I_s} \bar{x}_i, (s = 1, 2, \dots, H)$$



# SIR 알고리즘

[Step. 4] 각 분할  $s$ 에서 관측된 크기  $n_s$ 에 의해 가중된 설명변수  $X$ 의 분할평균에 대한 분산-공분산 행렬  $\widehat{\Sigma}_{X|y}$ 를 계산한다

$$\widehat{\Sigma}_{X|y} = \sum_{s=1}^H (1/n\hat{p}_s)(\hat{m}_s - \bar{x})(\hat{m}_s - \bar{x})^T$$

[Step. 5] 설명변수  $X$ 의 분산공분산 행렬  $\widehat{\Sigma}_X$ 를 계산한다

$$\widehat{\Sigma}_X = n^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$$



# SIR 알고리즘

[Step. 6] 다음의 일반화 고유값분해(generalized eigenvalue problem)를 수행한다

$$\widehat{\Sigma_{X|Y}} \hat{\eta}_i = \hat{\lambda}_i \hat{\Sigma}_X \hat{\eta}_i$$
$$\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p, i = 1, 2, \dots, p$$

위와 같은 알고리즘을 통해 추정되는  $\hat{\eta}$ 은 SIR의 충분 차원축소 방향이다.  
고유값의 크기순으로 순차적인 중요성을 지니게 되는  $i$ 번째 고유벡터  $\eta_i$ 는  $i$ 번째 SIR방향을 나타내며, 이때 상대적 크기가 큰 적절한 고유값에 의해  $d(\leq p)$ 개의 SIR방향을 선택한다.



# SIR 알고리즘 결과 해석

## Remark

- 표준화된 역회귀 곡선이 표준화된 충분축소차원 공간의 적절한 부분공간에 포함되는 경우, SIR은 모든 충분축소차원을 *recover* 할 수 없다.
- 예를들어 대칭함수  $g$ 가 존재할때,  $y = g(\beta_1 x) + \epsilon$ 에서  $\beta_1 x$ 가 0에 대칭이면,  $E(x|y) = 0$  이고,  $\hat{\beta}_1$ 은 추정의 의미가 없어진다.
- 이를 해결하는 방법은 모멘트를 더 높여서 추정하는 것이다.
- $x$ 가 정규분포라면, 어떤 방향으로든  $bx$ 는  $\beta_k x$ 들과 직교한다.
- 이때, 우리는  $Var(bx|y)$ 의 경우  $y$ 가 변하더라도 변하지 않는다는 것을 알 수 있다.



# Text Mining



# Text Mining

아래 식을 다항 역회귀 식으로 정의할 때,

$$x_i \sim MN(m_i, q_i) \text{ with } q_{ij} = \frac{\exp(\eta_{ij})}{\sum_{i=1}^p \exp(\eta_{ij})}, j = 1, \dots, p$$

여기서  $x_i$ 는 크기  $m_i$ 와 확률  $q_i$ 를 따르는  $p$ -차원 다항분포이고,  $\eta_{ij} = \alpha_j + u_{ij} + v_i^T \varphi_j$  이다. 이 때,  $v_i$ 를  $x_i$ 가  $y_i$ 에 따라 달라지는 반응 인자의  $K$ -벡터라고 가정했을 때, 선형 역회귀 식이 다음과 같다.

$$x_i = \Phi v_i + \epsilon_i$$

이 때,  $\Phi = [\varphi_1, \dots, \varphi_p]$ 는 역회귀 계수의  $p \times K (K \leq p)$ 행렬이고,  $\epsilon_i$ 는 오차항(크기  $p$ -벡터)



# 연속형 변수 $y$

## 단계함수

- 그 이후  $v_i$ 는 반응변수가 연속형일 때  $v_i = \text{step}(y_i)$ 라고하는 step function으로 본다  
한다.
- $y_i$ 를 step function에 넣고 조건에 맞으면  $y_i$ , 맞지 않으면 0을 취하는 것을 의미한다.

## 궁금증

- 이 때, 벡터  $\mathbf{v}$ 는 0 또는  $y_i$ 로 이루어진 것이 맞는가?
- 그렇다면  $y_i$ 의 조건을 어떤식으로 세워야 할까?
- 이 때, step function의 역할이 기울기를 활용하는 함수로 생각하고, step function 대신 시그모이드 함수  $\sigma(z) = \frac{1}{1+e^{-z}}$ 를 사용해도 되는가?
- 만약 그렇다면 연속형 변수라고 했을 때는 slice 갯수를 모든 data포인트의 수로 추정해야 할 것이다.(예상)





# Parameter 추정

## latent variable

- 지수 논문[3] 10쪽에서 나온

$$\pi(\varphi_j, \lambda_j | r, s) = \frac{\lambda_j}{2} e^{-\lambda_j |\varphi_j|} \frac{r^s}{\Gamma(s)} \lambda_j^{s-1} e^{-r \lambda_j}, s, r, \lambda_j > 0$$

에서 빨간 부분이  $\text{Laplace}(\varphi_j; \lambda)$ , 파란 부분이  $\lambda_j \sim \text{Gamma}(s, r)$ 이다.

- 이 때, 이 posterior 분포를 통해 모수를 추정할 때, Gibbs Sampling을 이용할 수 있을까. 여기서 r,s는 hyper parameter일 것이다.

## 궁금증(해결)

- 이 때 구한 추정된 모수( $\varphi$ )를 이용한다면 더 괜찮은 추정이 되지 않을까?
- Key논문(15페이지 4.Estimation 두번째 단락에 안되는 이유가 간략히 설명되어 있었음.)
- 2013년에 발표한 논문이라서 컴퓨터 성능이 예전보다 나아졌기 때문에 시도해볼만은 하지만, 현재 소유중인 컴퓨터로는 불가능할 것으로 판단(시도는 해볼 예정).



## Sliced inverse Regression



## 궁금한 점

### Question

- Inverse Regression의 방법을 단순 차원축소의 방법으로 사용하는지? -> 만약 그렇다면 SIR에 적용이 가능하다.

### 예상

- 차원 축소를 projection을 통해서 찾고 있다는 공통점.(이 때 유일한 조건은 조건부 기대치( $E[b^T X | \beta_1^T X]$ )가 선형이라는 조건)
- SIR 알고리즘에서도 slice할 때 Y의 범위에 따라 slice한다는 단순한 추가 알고리즘.
- Y의 범위를 일정하게 나누지 않아도 되고 slice 갯수와 성능은 상관없다는 논문의 해석이 있다.
- 그렇다면 smoothing spline 보다 구간을 나누는 것이 추정에 결정적이진 않으므로, 고차원 텍스트 데이터에 사용하기에 알맞을 수 있다.



# Slice 범위가 일정하지 않을 때

## Slice 범위가 일정하지 않을 때 필요조건

- 관측치의 동일한 분포를 보장하기 위해 각 슬라이스의 범위가 변화하는 경우는 다음과 같은 간격 선택을 따라야한다. 여기서  $F_y(\cdot)$ 은  $y$ 의 cdf.

$$I_h = (F_y^{-1}((s-1)/H), F_y^{-1}(s/H))$$



# Slice 수의 영향

## Slice 갯수가 중요하지 않는가?

that the following is one way to estimate  $E[\text{cov}(\mathbf{z} \mid y)]$ :

- (a) Introduce a large number of slices for partitioning the range of  $y$ .
- (b) Within each slice, form the sample covariance of  $\tilde{\mathbf{x}}_i$ 's that fall into that slice.
- (c) Form an average of the estimated conditional covariances of (b).

Intuitively, in order to get rid of the bias for estimating the conditional variance  $\text{cov}(\mathbf{z} \mid y)$  for each  $y$  in (b), we hope that the range of each slice will converge to 0, so that only local points will contribute to the estimation. But when the number of slices is too large, the sampling variance in each estimate of  $\text{cov}(\mathbf{z} \mid y)$  may not diminish, even for large  $n$ . Fortunately, the averaging process of (c) will stabilize the final estimate by the law of large numbers. As a matter of fact, even if the slice number is  $n/2$ , so that each slice contains only two observations, the resulting estimate will still be root  $n$  consistent.

*Remark 5.3.* Can SIR still yield reasonable estimates if the number of slices increases too fast and the number of observations in each slice is too small for  $\hat{m}_h$  to consistently estimate  $\mathbf{m}_h$ ? Remark 3.1 offers an answer. First, we see

The interesting connection of this estimate of  $E[\text{cov}(\mathbf{z} \mid y)]$  to SIR is that this estimate is proportional to  $I - \hat{V}$  because of the sample version of the identity given in Remark 3.1. Because of this conjugate relationship, a prin-



## slice와 공분산행렬의 관계

### Remark

결론 : 각 데이터 포인트들이 slice로 존재하여도 **WLLN**에 의해 공분산행렬의 평균은 안정화된다.

- 추가로 SIR에서  $E[cov(z|y)]$  추정량의 흥미로운 관계는  $I - \hat{V}$ 에 비례한다는 것
- 이 conjugate 관계때문에 최소  $K$ 주성분으로의  $E[cov(z|y)]$  추정량은  $\hat{V}$ 의 최대  $K$ 주성분과 같다.
- 이는 왜 slice의 수가 많아도 잘 작동되는지, 그리고 smoothing parameter의 선택만큼 결정적이지 않는지 보여준다.



## Future Study



# Future Study

## Future Study

- 이번 주에는 다시 한 번 논문들을 읽어보면서 생각해왔던 것들을 정리해 보았습니다.
- 지수논문에서 나온 데이터들을 이용해서 지수 논문의 방식대로 R에서 구현해볼 생각입니다.
- 아직 논리적인 수식전개부분이 Key논문에서 어려워서 막히는 부분이 있었는데, 지수논문을 다시 읽어보면서 해소할 수 있었습니다.
- 계속해서 논문을 읽어보면서 정리해보겠습니다.
- SIR에 대한 대략적인 모델 모형 설계가 끝나는대로 R 코드화 시켜보겠습니다.





# References



Cook, R. D., Li, B., **et al.** Dimension reduction for conditional mean in regression. **The Annals of Statistics** **30**, 455–474 (2002).



Eun Kyoung Seo, C. P. Effect of Dimension in Optimal Dimension Reduction Estimation for Conditional Mean Multivariate Regression. **Communications for Statistical Applications and Methods** **19**, 107–115 (2012).



Jisu Kim, L. 다항 역회귀와 한국어 리뷰 댓글을 이용한 영화 평점 예측. **41** (2020).



She, Y. **et al.** Thresholding-based iterative selection procedures for model selection and shrinkage. **Electronic Journal of statistics** **3**, 384–415 (2009).

